2010

Using Matching Methods From Both Fisher's Experimental Design and Rubin's Causal Model to Compare Between Two Medical Facilities with Extremely Skewed Number of Subjects

Gideon D. Bahn
Loyola University Chicago

LOYOLA UNIVERSITY CHICAGO

USING MATCHING METHODS FROM BOTH FISHER'S EXPERIMENTAL DESIGN AND RUBIN'S CAUSAL MODEL TO COMPARE BETWEEN TWO MEDICAL FACILITIES WITH AN EXTREMELY SKEWED NUMBER OF SUBJECTS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE GRADUATE SCHOOL

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

PROGRAM IN RESEARCH METHODOLOGY

BY

GIDEON D. BAHN

CHICAGO, IL

AUGUST 2010

ACKNOWLEDGEMENTS

I would like to thank all of the people who made this dissertation possible, starting with my wonderful professors in the Research Methodology Department at Loyola University Chicago. Dr. Teresa Pigott proved an excellent sounding board for me from the beginning of my time here, and steered me toward thinking critically in doing research and analyzing the real life data. Dr. Martha Wynne provided me with much perspective on issues of building survey questions and its methodology. Dr. Meng-Jia Bohanon also blessed me with her solid knowledge in multivariate analysis and feedbacks on this study. Finally, I would like to thank one of committee members, Dr. Kathleen Ruroede, who evoked me with the original data and its problem so as to probe the study questions from the beginning of this study and provided her keen insight to the end. All of their sage advice has put me on track when I veered precipitously away from my early goals, and their friendship and encouragement have made the difference in this long and arduous process.

I would also like to thank the Graduate School of Loyola University Chicago for providing the funds when the outside funding source was on the verge of discontinuation, which enabled me to continue and complete my research and writing.

My friends have provided me with a much needed cheering section. Some prayed for me so that I may be refreshed and ready to confront it all over again, and others, exchanging ideas for the study. In particular, I would like to thank Dr. Philip Hong at

Loyola University, Dr. John Jun in Chicago, David Kim, Jeremy Hajek and in Wheaton, Dr. Joe Shafer in the University of Penn State.

Finally, I would like to thank the love of my life and my best friend, Mary Bahn. Without her support and prodding, I would never have made it where I am today. Her unfailing good sense, great humor, and unparalleled companionship make me a very lucky man indeed.

Above all, I thank God my Savior, Jesus Christ who gave me physical and spiritual life, supported me with hope and love without ceasing, and to whom I dedicate all my work.

For my Savior, Jesus Christ, and my family: dad, Yehbyung; mom, JungHee; wife, Mary; daughter, Sarah; and son, Isaac.

There was no one who could stand an egg by itself except one person, Columbus.

Unknown

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

The present study deals with the problem of comparison between a two medical facilities' with extremely skewed sample sizes from non-experimental study. The data came from a study of rehabilitation interventions with patients diagnosed with cardiac and pulmonary issues who received treatment either in inpatient rehabilitation facilities (IRFs) or in skilled nursing facilities (SNFs). The main hypothesis was comparing the outcomes between the patients undergoing rehabilitation interventions administered at IRFs and the patients managed by SNFs. Due to inclusion and exclusion criteria, however, the study had failed to recruit sufficient number of participants between two comparison groups: 319 from IRFs and 27 from SNFs. As a result, the main hypothesis of the study was not tested due to the disparity of the participants between the two comparison groups, which could not be analyzed as a study with an unbalanced design because of lack of power in the analysis (Beacham, 2008).

In medical research, this kind of problem occurs often not only because of inclusion and exclusion criteria in recruiting patients for a study but also because of dropout patients due to many reasons, such as technical changes (certain insurance and/or Medicare policies eliminate possible participants), medical changes, or personal circumstances change in the middle of the study. By extracting *matching methods* from both Fisher's experimental design and Rubin's Causal Model (RCM) the present study attempts to offer ways to draw the causal inference in a non-experimental study with

sample size disparity between two comparison groups, especially when collected data disable a researcher to analyze.

The matched datasets were analyzed in two ways: multivariate of covariance (MANCOVA) first and two analysis of covariance (ANCOVA) models when there was a significant main effect in the previous MANCOVA model. No significant different effectiveness was found between IRFs and SNFs in the 1:1 Matched Data, but IRFs took better care than SNFs in the Caliper Matched Data, rehabilitating the patients diagnosed with cardiac and pulmonary diseases on the functional independent measure (FIM). In comparison methodology, the results suggested that both methods provided similarresults, but that Fisher's design fit better for small dataset while RCM, for larger data by using propensity scores to balance the matching sets.

CHAPTER ONE

INTRODUCTION

Experimental design has been a part of human existence since the beginning of time. To survive, human beings needed to find optimal places to hunt games or to plant crops to maximize their yield. In this way, humans were collecting data and analyzing it systematically to get information needed to make practical decisions long before the field of statistics existed. Modern-day research has developed in the same way, especially experimental research, by collecting complex data and turning it into useful information through analytic power. In many fields of scientific study, researchers plan and carry out systematic investigations of a phenomenon. When researchers are interested in cause and effect relationships, that is, when researchers wish to determine if the same action or event causes another to occur, they employ the principles of experimental design.

Experimental design has two fundamental elements: manipulation of independent variable (IV) and randomization. The manipulation of IVs should have a direct effect on dependent variables (DVs) in order to draw a causal inference (Holland, 1986). For example, a researcher plans to give two different medicines, A and B, to two separate groups of patients, ideally controlling all possible extraneous effects. However, some of the patients are already taking medication A or B. Here, when the researcher has no control over manipulation of the IVs—the participants are already taking the medicine

before the start of the experiment—only a correlation can be asserted between the IVs and the DVs. No causality can be drawn from the research when there is no manipulation of IV.

With manipulation of IVs, however, it is almost impossible to control all possible extraneous variables (variables other than the intended IVs) that can affect DVs. These uncontrolled-for extraneous variables are called confounding variables. The existence of confounding variables distorts the causal inference. In order to control the confounding variables, Fisher (1925) introduced the idea of randomization into research design. Randomization is achieved by random assignment of participants to experimental groups in order to control and quantify confounding effects on the outcome variables *a priori* and then measure the effects from the intended and manipulated treatment by the researcher (Levin, 1999).

In social, psychological, and medical research studies, however, sometimes manipulation of IVs and random assignment of participants are not feasible due to moral, ethical, and practical reasons. Therefore, the researchers have no control over statistically partialling out the confounding variables affecting the outcomes. When researchers have no control over manipulation of IVs and random assignment of participants, a non-experimental study is utilized. Due to the limitations in the research design in such non-experimental studies, causal effects are more difficult to infer. Especially in medical research, the random assignment of the participants is almost impossible when subjects are diagnosed with a certain disease or condition that requires treatment with a particular intervention, limiting the study to a quasi-experimental design. In addition, if a quasi-

experiment suffers from participant attrition, which often produces a small and/or different sample size between comparison groups, then the ability to draw causal inferences is compromised.

The present study deals with the problem of drawing causal inference from non-experimental studies with different sample sizes using a two-group comparison. The data came from a study of rehabilitation interventions with patients diagnosed with cardiac and pulmonary issues who received treatment either in inpatient rehabilitation facilities (IRFs) or in skilled nursing facilities (SNFs). The main hypothesis of the formal study was comparing the outcomes between patients with cardiac and pulmonary diagnoses undergoing rehabilitation interventions administered at IRFs and patients managed by SNFs. As a result of the inclusion and exclusion criteria, however, the study has failed to recruit sufficient number of participants between the two comparison groups: 319 from IRFs and 27 from SNFs. Due to the disparity in number between the two comparison groups, the main hypothesis of the formal study could not be analyzed as a study with an unbalanced design because of lack of power in the analysis; thus, it was not reported (Beacham, 2008).

In medical research, this kind of problem occurs quite often not only because of inclusion and exclusion criteria in recruiting the patients for a study but also because patients drop out in the middle of a study for a multitude of reasons. By extracting *matching methods* from both Fisher's experimental design and Rubin's Causal Model (RCM), the present study attempts to offer ways to draw a *causal inference* in a non-experimental study with sample-size disparity between two comparison groups,

particularly when the collected data limits a researcher to analysis with unconventional methods.

The fundamental theory of both methods is *matching* background characteristics of the participating patients in order to draw *causal inference* between the two comparison groups. This study hypothesizes that the matched sets from both Fisher's design and RCM will provide similar results with the same type of analysis. Though this study only utilizes *matching methods* from two designs, the theoretical underpinnings of the methods will be reviewed in order for the readers to understand why the proposed methodological interventions are employed in this study.

CHAPTER TWO

REVIEW OF RELATED LITERATURE

In order to draw *causal inferences* in a research study, a researcher has to demonstrate the relationship between the treatment (x) and the outcome (y) in two ways; first, that x causes y, and second, that y does not occur without x. This is the foundation of the "gold standard" study in experimental design (Trochim, 2006). In a gold standard study, a completely randomized design is used to establish two groups that are comparable so that a causal relationship between the treatment and the outcome can be verified. For example, imagine that a researcher wants to determine the effect of a medicine on a headache. If a person with a headache takes a medicine and the headache disappears, this is insufficient to draw a causal relationship between x (medicine) and y (no headache). In order to establish the second point, another person with a headache who takes only a placebo (C) should be included. If the person who takes C and continues to have a headache, can the researcher draw a conclusion concerning the causal effect between x and y? What if y appeared not due to x but due to some unknown variable? What if one person who takes x is healthier and younger than another person taking the placebo? The internal validity problem arising from using only two people to establish a causal effect can be addressed by increasing the number of participants in both groups (*replication*). In this way, it can be established that y did not occur by an unknown

variable, rather it systematically occurred by x. Yet, merely adding to the number of participants without experimental control can bring additional problems caused by potential differences between participants, such as age, gender, health conditions, et al. In order to control confounding characteristics of the participants from systematically influencing the outcome (y), a researcher could *randomly* assign participants to either one of the groups in the study (*randomization*). R.A. Fisher (1926) is considered to be the one who conceptualized a systematic solution to control confounding characteristics though *replication, randomization, blocking* and *matching* in *experimental design*.

Experimental Design by R. A. Fisher

In this example above, two very important concepts are introduced for an experimental design: *replication* and *randomization*. One person is not sufficient to draw a causal relationship between x and y from a study; rather, *replication* and *randomization* are needed. Through *replication*, a researcher is able to estimate more valid and precise causal effects (Kuehl, 2000). Through randomization of the participants, a researcher is able to control any possible characteristics influencing the outcome (y) except for the treatment (x). R. A. Fisher built this theory of *replication* and *randomization*. Further, he expanded the theory of *replication* and *randomization* to include *blocking* and *matching* in experimental design.

Replication and Randomization

Simply speaking, *replication* is the addition of more participants (subjects or fields) in each comparison group of a study as in the previous example Kuehl (1994, p. 14) asserts,

"The scientific community regards replication of experiments to be a prime requisite for valid experimental results." *Replication* lays the foundation for *randomization* and *randomized block analytical paradigms* (Fisher, 1926). For example, if there is only one subject in the treatment group and another in the control group, a researcher cannot be certain that these two subjects can provide a representative and accurate measure of the treatment effect. Including more subjects in the study increases the validity and precision of the causal effect (Fisher, 1925). With more participants, the researcher can increase the confidence level that the outcome is a "typical" response to the treatment.

Many researchers have applied and tried to expand the benefits of replication, but Kuehl (1994, p. 14) provides one of the best explanations:

- ★ Replication demonstrates the results to be reproducible, at least under the current experimental conditions.
- ★ Replication provides a degree of insurance against aberrant results in the experiment due to unforeseen accidents.
- ★ Replication provides the means to estimate experimental error variance. Even if prior experimentation provided estimates of variance, the estimate from the present experiment may be more accurate because it reflects the current behavior of observations.
- ★ Replication provides the capacity to increase the precision for estimates of treatment means.

Even with r*eplication,* the validity of causal effects is limited when there are known differences between the two comparison groups. For example, if the participants in the treatment group are younger than those in the control group, the responses may be influenced not only by the treatment but also by the age difference. If most participants in the treatment are older women with health concerns, then their responses may be influenced by these facts more so than by the treatment. In this case, the differences between the responses in the treatment and the control groups cannot be attributed to the

causal effect of the treatment. The researcher faces a dilemma of how to assign the participants into two groups in order to control any differences between groups at the outset of the experiment. Fisher (1926) asserted that *randomization* is one of solutions to the dilemma of systematic group differences in an experimental design.

Randomization is randomly assigning participants into two comparison groups. The randomization method controls any known and especially unknown extraneous variable(s) that can influence the effect of the treatment on the outcomes. Considering the unknown extraneous variables, how much difference between two groups or similarity among members of one group should be factored into the analysis? By randomly assigning participants into two comparison groups, any possible, but unknown extraneous variables are controlled. Through this random assignment, any personal preference and/or unknown characteristics of participants are controlled from the outset. Fisher (1926) remarked, "One way of making sure that a valid estimate of error will be obtained is to arrange the plots deliberately at random, so that no distinction can creep in between pairs of plots treated alike and pairs treated differently….." (p. 506-507) Therefore, a researcher is able to ensure the equality of two comparison groups through randomization.

In the presence of known background differences between the two comparison groups, however, randomization is not sufficient to control any possible confounding variable that systematically affects the outcomes. In this case, *blocking* and *matching* methods could be used. The idea of randomization does not always guarantee a valid causal inference, especially when the study is poorly designed (Kuehl, 1994). When there

are known *a priori* differences among participants, *blocking* and *matching* methods enable the researcher to control their unwelcomed effects in a research study.

Blocking

Blocking is grouping participants with known characteristics in order to control and study the relationship between the independent variable and the dependent variable by controlling for the extraneous one. Fisher (1925) suggested *blocking* to control known confounding variables due to the heterogeneity of participants. By blocking participants into homogeneous groups, a researcher reduces experimental error. Blocks are often used as one of the independent variables in addition to the treatment effect; such as age in the example below.

Table 1

Block Design

	Younger age group C1	Middle age group C2	Older age group C3
Treatment A	Block 1	Block 2	Block 3
Treatment B	Block 4	Block 5	Block 6

Randomized complete block design (RCB) is introduced as the exemplary block design by Fisher (1926). For two comparison treatments, RCB randomly distributes the participants with known characteristics proportionately to each block. For example, suppose a researcher is interested in comparing two treatments, treatment A and treatment B, among three different age groups, consisting of younger (15-30 year old),

middle aged (31-55 year old) and older (55 and older), because a former study suggested different outcomes among age group levels. Since there are two treatments and three age groups there are six blocks as in Table 1.

All the participants are distributed randomly in 6 blocks. In this way, a researcher includes all the treatments in each block and participants are randomly assigned to the treatments. It is noted that an equal number of participants should be distributed to each block in RCB. Fisher (1925) showed that the experimental error was reduced about 55% by utilizing the random block experimental design. The difference between age groups is one of the researcher's interests in this block design.

Since Fisher's pioneering work, many new block designs have been developed. Balanced incomplete block design (BIB) is as precise as RCB with a fewer number of blocks. In the BIB family, there are three types of designs: row-column designs, partially balanced block design (PBIB) and unbalanced incomplete block design (UIBD). Row-column design was developed for two blocks. PBIB has an advantage because it reduces replications, and UIBD was developed for unequal replication in different blocks when comparing two or more groups.

Blocking methods can be utilized to control known confounding variables in many fields of study. In medical studies, randomized block designs can control for any known characteristics between the treatment and the control groups, such as age, weight, and other physical characteristics. When randomization and blocking method are utilized blindly, however, the resulting comparison groups may be different. For example, there is a small number of people who have been diagnosed with diabetes among all participants.

If a researcher randomly assigns all the participants in two comparison groups without considering this known special group of participants, the researcher may assign too few or too many of the diabetics into one group. Although a researcher strived to achieve equality of the comparison groups by randomly assigning participants into blocks, the groups may not be comparable because of existing different background characteristics. To address this problem, *matching methods* provide one solution.

Matching Methods

Matching methods are used to randomly assign participants into groups based on their known background characteristics. As in the example above, a researcher may assign the same number of participants with diabetes between the two comparison groups. Suppose when twelve participants with diabetes are found; two in younger group, four in middle aged group and four in older group, a researcher has to assign them equally to each treatment. This means that one participant with diabetes in treatment A and another in the treatment B in younger group, and two in the treatment A and another two in the treatment B in middle aged group and so on. In this way, the matching method is utilized to control known variables, which can systematically influence the outcome variables (DVs). Thus, through matching, the researcher establishes comparability between groups with known background characteristics that are often called *covariates*. In general, the matching strategy has two branches; *pair matching* and *non-pair matching* (Kuehl, 1994).

The *pair matching* strategy creates matching pairs on known background characteristics between participants into two comparison groups (Kuehl, 2000). When they are paired one to one, it is called *1:1 matching*. However, exact 1:1 matching is often not feasible, especially with continuous variables because it is almost impossible to match the exact value. For example, when there is a thirty-year-old participant with diabetes in one group, a researcher may not be able to match a participant with diabetes with the same age in another. In that case, a researcher has to match groups of similar values through a technique called *caliper matching*. In caliper matching, the same number of participants in an age group (i.e. younger, middle and older) between comparison groups, treatment A and B, is matched.

In *non-pair matching*, two strategies are developed: one with *frequency counts* and another with *mean values*. Both strategies were established in order to control the known covariates. The *frequency* approach stratifies participants based on the frequency distribution of a covariate and matches them between two comparison groups so that there are sufficient numbers of participants in each stratum. For example, suppose there are no participants in one group between 200 and 220 pounds but five between 220 and 225 in body weight. In order to have a sufficient number of participants, a researcher has to make an interval 220-225 pounds as a stratum, so that both comparison groups have a sufficient number of participants, at least five in this example. The *mean* approach also stratifies participants based on the mean of a covariate such that each stratum between comparison groups has a similar mean of the covariate. A similar mean weight of a

certain number of participants in each stratum between two comparison groups should be used to match them in the example above.

In many fields of study, such as sociology, psychology and medicine, an experimental design is not feasible due to moral, ethical and practical reasons. In this case, a researcher is left with *non-randomized experimental designs* that are both *observational* and *quasi-experimental*. Traditionally, non-randomized experiments do not make causal conclusions between treatments and resulting outcomes because there are potential confounding variables, which will distort the degree and direction of the estimated effects of treatment on subsequent outcomes. As mentioned above, in *non-experimental* designs, confounding effects occur because the participants are not randomly assigned by researchers making the groups suspect for systematic differences.

Observational studies have neither control over the manipulation of the treatment assignment of independent variables (IVs) nor random assignment of the participants. Some studies in the medical field can be observational. A researcher may observe different records of two hospitals in order to compare two different ways of taking care of patients: a new way versus an old way or two different ways, A or B. Patients in these hospitals are neither randomly assigned nor received a different treatment designed by a researcher who assumes that both hospital received patients who have similar background characteristics. But in most cases, the patients have differences in demographics and medical history. As a result, unknown extraneous variables can systematically influence the difference in the outcome variable of interest the researcher is intending to compare between the two hospitals.

Quasi-experimental studies have control over the treatment assignment but not the random assignment of participants. Using the example above, in the case of quasi-experimental design the researcher can have a group of patients who are willing to participate in an experiment among the patients admitted to the two hospitals, and apply the treatments the research designed for a special group of patients. The lack of random assignment, however, limits a quasi-experimental design to draw causal inference between the treatments (IVs) and the responses (DVs).

To overcome the limitations in non-experimental studies, both observational and quasi-experimental, many researchers have tried to draw causal conclusions. Cochran and Rubin (1973) attempted to make causal inference in observational data using matching and blocking, and Rubin (1973, 1974) established Rubin's causal model (RCM) in *non-experimental design*.

Rubin's Causal Model

Rubin (1973, 1974) began developing the theory for causal inference application from *non-experimental* studies. Holland (1986) named this idea Rubin's Model and Rubin (2004) himself, called it Rubin's Causal Model (RCM). RCM is approached within the *potential outcome* theory from Neyman (Rubin, 2004), and the philosophical background of RCM stems from Hume (Holland, 1989).

The Neyman's theory estimates a causal effect using both *observed* outcomes and *unobserved* (*potential*) outcomes in an experiment (Neyman, 1923; Dabrowska and Speed, 1900). In order to find a causal effect of one treatment comparison to another in

an experiment, a researcher usually establishes two groups, a treatment group and a control group. Participants are randomly assigned to either the treatment group (A) or the control group (B). This assumes that each participant also has the *potential* to be assigned to a different group and receive a particular treatment, A or B. The researcher observes two outcomes from the two different treatments, respectively, and compares them in order to draw causal inference. This scheme, however, is a modified way to find the true causal inference estimate. Based on Hume (1740, 1748), the true causal inference is the difference between the outcomes of participants when they are both in the treatment and in the control group. This is the philosophical foundation of RCM (Holland, 1989).

In order to follow Hume's theory to draw causal inference, each person should be assigned to both treatments; however, it is impossible to assign the same person to two different groups, treatment and control, concurrently. The only way to do so is to carry out two experimental studies with same design; one after another with enough time to wash out the effect of the first experiment. This is unrealistic however, for economic, ethical and practical reasons, especially in medical and psychological field of study. With only one experiment, Neyman (Neyman, 1923; Dabrowska and Speed, 1900) proposed to estimate the *potential outcomes* in order to draw causal inference of the treatment effect. Rubin (Holland, 1989) developed this idea into the statistical model, called RCM.

Therefore, in an experimental study there are two *observed* outcomes and two *unobserved* outcomes within RCM. In Table 2, the *observed* outcome from the treatment group is denoted as Y(t), and the *observed* outcome from the control group, Y(c). The *unobserved* outcome from the treatment group is denoted as Y(t)*, and the *unobserved*

outcome from the control group, Y(c)*, which are not measured. Rubin (1974, 1977) proposed to estimate the *unobserved* outcomes based on the *observed* outcomes. The estimated outcomes are called "*potential*" outcomes.

Table 2

Rubin's Theory

	Yt*i*=potential outcomes	Yc*i*=potential outcomes
Treatment	Observed outcome=Y(t)	Not observed=Y(t)*
Control	Not observed=Y(c)*	Observed outcome=Y(c)

For example, suppose 200 participants are randomly assigned into two comparison groups: 100 in the treatment group and another 100 in the control group. The medicine (t) is given to the treatment group, and the placebo (c) to the control group. Two responses are *observed*, Y(t) from the treatment group and Y(c) from the control group in Table 2. The outcomes; however, Y(t)* and Y(c)* are *not observed*. The *potential* outcome, Y(t)*, is impossible to measure because the participants who are in the treatment group cannot receive a placebo at the same time. Y(c)*, another unobserved outcome in the control group, is also a *potential* outcome of the participants who cannot go back to the beginning of the experiment to receive the treatment. The *potential* causal effects are supposedly $Yt_i=\sum(Y(t_i)-Y(c_i)^*)$ and $Yc_i=\sum(Y(t_i)^*-Y(c_i))$, where i=1-100, however, $Y(c_i)^*$ and $Y(t_i)^*$ are not observed but estimated as explained above. Therefore, the average of the two *potential* causal effects is $Y_i=(Yt_i+Yc_i)/2$, which is called "*the average causal effect*" in RCM (Rubin, 2000).

Based on the RCM, many methods have been developed to draw causal inference in non-experimental studies. Even though non-experimental studies lack manipulation of the treatments and random assignment of the participants, Rubin (1974) suggested that a researcher may be able to establish the comparability of two groups when participants of each group are matched based on observed background characteristics. He claims, "The basic conclusion is that randomization should be employed whenever possible, but that the use of carefully controlled nonrandomized data to estimate causal effects is a reasonable…." (p. 688). But matching participants between two comparison groups reaches a limitation with many background variables, especially with continuous variables such as age and blood pressure levels. Rosenbaum and Rubin (1983) advanced the idea of *propensity scores* in order to accommodate many covariates, including categorical and continuous variables.

Propensity scores

A *propensity score* is a probability of being in an assigned group (either treatment or control in the example above) that is calculated based on the similar background characteristics of the participants collected before the experiment, called *covariates*. Rosenbaum and Rubin (1983) defined the *propensity score* as "the *conditional probability* of assignment to a particular treatment given a vector of observed covariates," (p 41). The *propensity* scores can be written as $P(t_j|A_k)$, where t_j = the treatment group ($_j$ =1) or the control group ($_j$ =2), A_k=the number of the covariates (k=1, 2, 3, ….., *k*). The *propensity scores* are known in experimental studies by random assignment, which

should not be significantly different between the two groups. Through randomization, the causal effect is drawn through the direct comparison between the treatment group and the control group with the same covariates. Therefore, the propensity scores are the function of covariates, A_k. In non-experimental studies, however, the propensity scores are unknown, yet they can be estimated in maximum likelihood logistic regression (Hosmer & Lemeshow, 2003) with the vector of covariates,

$A_k(t_j)=A_k(t_1)+A_k(t_2)$, where $A_k(t_1)=A1(1)$, $A2(1),\ldots,Ak(1)$ for the treatment group and $A_k(t_2)=A1(2)$, $A2(2),\ldots,Ak(2)$ for the control group.

$\text{Log}(\pi_{jk}/(1-\pi_{jk}))= A_k(t_j)\gamma_{jk}$, where γ_{jk}=a vector of the number of coefficients and t_j=(t=1, c=2).

For *individual propensity score*: $\pi_{jk}=e\text{^}(A_k(t_j)\ \gamma_{jk})/(1+ e\text{^}(A_k(t_j)\ \gamma_{jk}))$ (3.2.1)

Therefore, *propensity score*, $P(t_j|A_k)= \pi_{jk.}$

Rosenbaum and Rubin (1983) asserted that the *propensity score* can be used as a *balancing score* between two comparison groups based on covariates in a non-experimental study, assuming the study has "strongly ignorable treatment assignment." The essential idea of using *propensity scores* comes from equating two or more groups of participants in a study by matching and blocking data obtained from the groups in *nonexperimental design studies*. Even careless randomization in a study with experimental design may bring unwanted confounding variables that will affect the responses (DVs), especially when there are unintended similar background characteristics among the participants in only one of the groups (Kuehl, 1994). Although a non-randomization study lacks manipulation of the treatment and random assignment, if all

possible background characteristics are matched in all levels of independent variables (IVs)—called "strongly ignorable treatment assignment"—it is reasonable to draw causal inference through the observation of the responses (Rubin, 1975). The problem potentially remains with collecting all "*possible*" background characteristics when there are only known or "*observed"* covariates.

Rosenbaum and Rubin (1983, p. 41) expanded the parameters of the *propensity score*:

> ….. Both large and small sample theory show that adjustment for the scalar propensity score is sufficient to remove bias due to all observed covariates. Applications include: (i) matched sampling on the univariate propensity score, which is a generalization of discriminant matching; (ii) multivariate adjustment by subclassification on the propensity score, where the same subclasses are used to estimate treatment effects for all outcome variables and in all subpopulations; and (iii) visual representation of multivariate covariance adjustment by a two dimensional plot.

Since then, many people have developed the precision of the individual propensity scores in robustness (Albert & Chib, 1993), in robit model (Liu, 2004), in semi-parametric (Breiman et al., 1984; Luellen, Shadish & Clark, 2005), and in neural networks (King & Zeng, 2002).

These *individual propensity scores* are valuable in many ways. Rosenbaum & Rubin (1983) proposed to use the propensity scores as *balancing scores* in matching participants between the treatment group and the control group. The average causal effect, $Y_i=(Yt_i+Yc_i)/2$, is an unbiased estimate because it is adjusted by the propensity scores (Rosenbaum and Rubin, 1983). The propensity scores are ultimately used to match the participants between the two comparison groups based on "*observed"* covariates that

meet the assumption of “strongly ignorable treatment assignment” and draw causal inference in nonexperimental studies. Rosenbaum and Rubin (1983, 1984) utilized the propensity scores extensively in the *matching method* and *subclassification* to reduce any bias due to experimental error.

Matching Method and Subclassification

The idea in *matching method* and *subclasscification* using propensity scores in RCM is stemmed from Fisher’s matching and blocking in the experimental design. The *matching method* is employed in RCM using propensity scores in order to reduce possible bias (errors) that affect the responses. The matching method in RCM utilizes propensity scores to balance participants between the two groups based on observed covariates. *Subclassification* is used to establish a few subgroups of participants between comparison groups based on the levels of the propensity scores similar to establishing blocks in an experimental study.

Many researchers have employed the idea of *matching methods* using propensity scores in nonexperimental studies. By using individual propensity scores, participants are *matched* between the two comparison groups in order to have similar distribution of the covariates (Rosenbaum, 2002). Though *1:1 matching* is an ideal matched set, matching exact propensity scores from one group to another is almost impossible in many cases because the propensity scores are continuous number such as weight and height. A C*aliper matching*, which includes matching few closer scores, is more feasible than an exact matching (Schafer and Kang, 2008). *Full matching method*, involves using

propensity scores of all participants within subclasses including at least one from either comparison groups, and is another alternative way of matching (Rosenbaum, 1991; Hansen, 2004).

Propensity scores, however, do *not always* perfectly balance two comparison groups with the covariates, especially when there are multiple covariates. In order to achieve the *assumption* of "strongly ignorable treatment assignment" using propensity scores, it is critical to as balancing scores, collect as many covariates as possible. Many covariates, however, often cause difficulty in matching with mathematical adjustment due to different levels of propensity scores. As mentioned above, the propensity scores are probabilities between 0 and 1. For example, one subgroup of participants within the treatment group has smaller propensity scores between 0.1 and 0.4, and another subgroup has larger propensity scores between 0.7 and 0.9, while participants in the control group are spread widely between 0.3 and 0.9. If all participants between the treatment and the control group are compared after matching them with the propensity scores, the estimated causal effect is inappropriate because two groups essentially have different levels of propensity scores: the control group include a subgroup with propensity scores between 0.4 and 0.7. This problem is called *extrapolation*. Rosenbaum and Rubin (1983) advanced and expanded the idea of *subclassification* in order to resolve this problem.

Subclassification is dividing subjects into several subclasses of people based on the level of propensity scores between the two comparison groups. The main idea is that the participants in the treatment group who have a lower level of propensity scores should be compared with the participant in the control group with the same level of propensity

scores. Based on the example above, a researcher can make *subclasses* with similar levels of propensity scores; *one matched subclass* with a lower range between 0.3 and 0.4 and *another subclass* with a range between 0.7 and 0.9. *Subclassification* can be viewed analogous to matching within a block design in the experimental study. Often the ranges of propensity scores have different percentile levels. Through matching within subclasses according to the percentiles of propensity scores, more homogeneous blocks are created for comparison. Cochran (1968) showed that *subclassfication* in univariate analysis removed 90% of bias caused by covariates in observational studies. Rosenbaum and Rubin (1984) also proved that five subclasses based on the levels of propensity scores reduced over 90% of the bias caused by the covariates.

Matching method using propensity scores is more convenient when there are many covariates in matching participants between two comparison groups than Fisher's "hand-matching". When there is a wide range of propensity scores between 0 and 1, however, *subclassification* is useful to match with subgroups of people based on different levels of propensity scores. The present study will utilize two *matching, one-to-one and caliper,* from both method, Fisher's and RCM. *Full* matching and *subclassification* in RCM are not used because there is no comparable datasets from Fisher's method.

CHAPTER THREE

METHODOLOGY

Introduction

The present study deals with data where there is a discrepancy in the number of participants between two comparison groups after the pre and posttest scores were collected. Due to the extreme discrepancy of the number of participants between the two types of facilities, no final analysis concerning the formal research question was reported (Beacham, 2008). Such data limits a researcher in finding a comparative effectiveness between two groups.

The purpose of this study is to resolve the analytical problem of extremely skewed numbers of participants between two comparison groups. The best resolution would be to recruit a sufficient number of participants for both groups so that the comparison is reasonable. But, in many cases, it is not possible due to ethical, medical, and/or practical reasons, especially after the data collection stage is over. In order to resolve this analytical problem, this study proposes using two data matching methods: matching methods using Rubin's Causal Model (RCM) and matching methods in Fisher's experimental design. Rosenbaum and Rubin (1983) proposed matching participants between two groups using propensity scores in non-experimental data. The propensity scores are calculated based on a set of covariates (CVs), which are the background

characteristics of the participant before he or she receives a treatment. In theory, the "propensity matching method" is applicable to both small and large datasets according to Rosenbaum and Rubin (1983), but the method has been developed extensively only for a large number of subjects. This study will apply the propensity matching method to a dataset with a small sample.

Fisher (1926) suggested using a matching method in an experimental design in order to establish comparable groups and control any possible bias. This study also proposes utilizing Fisher's matching method with non-experimental data, importing Fisher's idea of matching two comparable groups through their background characteristics before a treatment. Fisher's matching method matches through raw scores of background characteristics, while the RCM matching method matches participants through propensity scores. The matching method employed from Fisher's experimental design is called "*hand matching*," and the one employed from RCM is called *propensity matching*.

Therefore, this study hypothesizes that matched datasets using the two proposed matching methods should produce similar results with the same analysis. The specific hypotheses guiding the study are:

1. What are the differences, if any, in the results after analysis of the matched datasets produced by the two matching methods, Fisher's and RCM?
2. What are the similarities, if any, in the results after analysis of the matched datasets produced by the two matching methods, Fisher's and RCM?

3. What adjustments could be made in the matching methods to control for the differences found in i), if possible?

In order to illustrate the methodological intervention, a part of the data in the cardiac-pulmonary database described below will be utilized.

Data Collection

The present study uses data collected by research team members and supporting staff from eight facilities funded by the American Medical Providers Rehabilitation Association (AMPRA). The patients, who are diagnosed with cardiac and/or pulmonary diseases, are usually discharged from an acute care system to rehabilitation facilities. A major number of patients are discharged to either inpatient rehabilitation facilities (IRFs) or skilled nursing facilities (SNFs). The formal study was comparing these two types of facilities as an independent variable (IV) in rehabilitating the patients diagnosed with cardiac and/or pulmonary diseases.

The treatment at both IRFs and SNFs was defined as the "usual care" of each facility in rehabilitation therapy regimens for the study patients; given medical policies and financial reimbursement systems for each level of care. After each patient agreed to participate in the study and before starting the treatment study, therapist and nurses collected each participant's background characteristics such as demographics and pretest scores as covariates (CVs). After the treatment and at the time of discharge, the assigned therapist and nurses also collected posttest scores as dependent variables (DVs). At the

admission and discharge, four domains were measured: 1) medical/physiological, 2) functional, 3) psychosocial, and 4) behavioral.

For this present study, however, not all four domains were utilized in matching the datasets: only demographics and some pretest scores were used as background characteristics as seen in Table 3. The ten CVs used to match participants in this study were primary and secondary diagnosis, pretest scores on the Charlson comorbidity index (ChCom), the stress/anxiety index (StAnA), Funtional Independent Measure (FIMA), SF-12 Health Survey (SF12A), Health Care Utilization (MUtil), and demographics (age, weight, and gender).

Table 3

Pre- and Posttest Data

Study Components	Pretest-admission (CVs)	Posttest-discharge (DVs)
Physical/Functional Assessment	Functional Independence Measure (FIMA)	Functional Independent Measure (FIMD)
Psychological Assessment	Quality of life (SF12A) Health Care Utilization (MUtil) Stress/Anxiety Index (StAnA)	Quality of life (SF12D)
Medical and Physiological	Demographic variables Charlson's Co-morbidity Index (ChCom)	

Two posttest scores—Functional Independence Measure (FIMD) and SF-12 Health Survey (SF12D)—were considered as DVs. The Functional Independence Measure (FIM) measures the severity of disability using 18 items on a 7-Likert scale: 13 motor items and five cognitive items (Mackintosh, 2008). FIM (either pre or posttest) is usually measured by study therapists, nurses and clinician's accessment. It is a measure of how independent a patient is based on18 items. SF-12 is a twelve-question version of

the Medical Outcomes Study Short-Form 36-Item Survey (King et al., 2009). SF-12 provides two measures: physical and mental health. It can be completed by the patient or by study nurses questioning patients directly. So, the presence of the patient is required to fill out the survey.

In order to ensure diversity in the population of the study, the participants were recruited from eight different geographic locations with a total target sample size of 1,200: 800 from IRFs and 400 from SNFs. Before the study, the participants were assigned in one or the other facility by the physicians, and approved by insurance (certain insurance companies do limit a patient's discharge to IRFs for cardiac and pulmonary rehabilitation). This fact limited the research study design to a non-randomization method. Then patients were identified for the study based on inclusion and exclusion criteria in each facility. The inclusion rules were 1) the ability to follow written and oral instructions in English; 2) equal or older than 21 years of age 3) the ability to tolerate three hours of total therapy per day; 4) to have a plan for return home or other community destination; 5) to have an adequate support system; and 6) to be medically stable as indicated by either blood pressure, heart rhythm, heart rate, and afebrile status, or by improving blood count and chemistries. The exclusion rules were 1) refusal to participate in the study, 2) inability to follow written or oral instructions, 3) inability to speak or write English, 4) under 21 years of age, and 5) ventilator dependent (Skolnick, 2008). However, after screening the potential participants using the inclusion and exclusion criteria, the final number of participants for the study only reached 346: 27 from SNFs and 319 from IRFs. Due to the sample size disparity, which limited power in the analysis,

the main research hypothesis, comparing IRFs and SNFs, was not conducted in the first report (Beacham, 2008).

In many research fields, it is often impossible to design an experimental study due to medical, ethical, and moral constraints. Moreover, the participants who were recruited can drop out from the study for unexpected reasons, including technical changes (certain insurance and/or Medicare policy restrictions eliminating possible participants), medical changes, and/or personal circumstances. This often makes a researcher unable to analyze the data and draw inferences due to the difference in number of participants and data between two comparison groups, resulting in a lack of analytical power.

By importing *matching methods* from both Fisher and RCM, the present study attempts to offer ways to draw inferences when the sample size disparity makes it impossible to analyze data through conventional approaches. Based on Rosenbaum and Rubin (1983), the *propensity matching method* should reduce bias, just as in a completely randomized design, with large or small datasets. The *hand matching method* in Fisher's experimental design is utilized in this study in order to establish comparable groups based on known background characteristics (covariates) among the participants in non-experimental data.

Matching Methods

This section describes how the two matching methods, Fisher's and RCM, were implemented in this study. Fisher's matching method has two matching procedures, one-to-one (*1:1*) and *caliper* matching, while RCM matching method has four: *1:1*, *caliper*,

full and *subclassification* matching. From RCM matching method, however, only two matching methods will be utilized—*1:1* and *caliper*—because there is no dataset from Fisher's method to compare with datasets made through *full* and *subclassification* matching in the RCM method. In this study, the pretest scores and demographics of each patient will be used as the background characteristics (CVs) for matching.

Table 4

Primary and Secondary Diagnosis

Cardiac diagnosis*	IRFs	SNFs	Pulmonary diagnosis*	IRFs	SNFs
Coronary Artery Bypass#	77	3	Chronic Obstructive#	166	7
Valve-Replacement#	22	4	Inter#	2	
Myocardial Infarction#	13	1	Pre/Post Toraic#	2	
Congestive heart failure#	9	12	Other	5	
Ventricular Assistive Device#	6				
Cardiac Transplantation#	1				
Pre/post Thoracic Surgery#	4				
Other#	11				
Total	143	20		175	7

* = Primary diagnosis, # = Secondary diagnosis

Two matching methods, Fisher's "*hand-matching*" and RCM *propensity matching*, are employed in order to establish comparable datasets between IRFs and SNFs given different sample sizes. First, the plan of "*hand-matching*" will be reviewed in the order of *1:1* and *caliper matching* procedures. Second, the plan of *propensity matching* will be explained in the same order: *1:1* and *caliper matching*.

Both strategies of matching in this study starts with the primary diagnoses, cardiac and pulmonary, between the two groups of patients in Table 3.2 below. In each primary diagnosis, there are secondary diagnoses (subcategories): eight in cardiac and four in pulmonary. For *hand-matching*, there will be priority among the CVs, as well as the

primary and secondary diagnosis in matching the participants between the two facilities. The matching will be done by visually screening raw scores. *Propensity matching* will treat the pretest scores, demographic variables, and primary diagnoses equally as covariates in matching. Each plan is described in detail below.

Hand-matching from Fisher's experimental design

Hand-matching is a matching method imported from Fisher's experimental design for the present study with *non-experimental* data. Due to matching the complex data in this study with many CVs, including primary and secondary diagnosis, it is impossible to match the participants with all CVs with the present data. Therefore, a priority among CVs will be given in matching after consulting two experts in the area of health and rehabilitation care. The order of the priority among CVs are as follow: 1) primary diagnosis and 2) secondary diagnosis: and pretest scores of 3) Functional Independent Measure (FIMA), 4) SF-12 Health Survey (SF12A), 5) Health Care Utilization (MUtil), 6) Charlson comorbidility index (ChCom), 7) the perceived stress index (StAnA), 8) age, 9) weight, and 10) gender. With this priority among CVs, 27 participants from SNFs will be matched with 27 participants from IRFs. For example, first, three SNF patients who were diagnosed with coronary artery bypass (secondary diagnosis) within a cardiac diagnosis (primary) will be matched with another three out of 77 IRF patients with the same diagnosis in Table 4. Then, in the order of priority of the other CVs, the raw values of the pretest scores and demographics (weight, gender, and age), will be used to match.

According to the literature review above, there are two possible *hand-matching* methods: *pair* and *non-pair matching*. Comparing between IRFs and SNFs, the *pair* matching method will be utilized because it makes sense to pair patients between two comparison groups. In *pair* matching, there are two procedures: *1:1* and *caliper matching*.

The *1:1 matching* is also called *exact* matching because the samples are matched using exact values of covariates. With a small number of subjects, it may not be feasible to match with the exact values between two groups. When an exact value is not found for matching, the closest value will be matched. For example, when one 45-year-old male patient from SNFs cannot be matched to a 45-year-old male patient from IRFs, a male closest in age from IRFs will be matched.

However, when not enough cases are available to match within the *secondary* diagnosis between two groups, unmatched participants from other secondary diagnoses will be used to match. For example, in all *secondary* diagnoses, the number of patients from IRFs is greater than that of SNFs except in *congestive heart failure* in Table 3.2. Since there are only 9 patients with *congestive heart failure* in IRFs, there are not enough cases to match all 12 patients from SNFs. The number of possible matches will be reduced from 27 to 24 when the participants should be matched within *secondary* diagnosis. Therefore, the unmatched participants who were diagnosed with all *secondary* diagnoses from IRFs will be matched to a participant from SNFs with *congestive heart failure* under *cardiac* diagnosis in Table 3.2. In this way, 27 participants from SNFs will be able to be matched with 27 from IRFs.

In *caliper* matching, participants will be matched in the same manner of *1:1 matching* using the same priority among CVs in the ratio of k:1 between IRFs and SNFs. For example, with a 3:1 ratio, three patients from IRFs are matched to a single patient from SNFs until all participants from SNFs are matched. This means 81 participants from IRFs will be matched to 27 from SNFs.

Propensity Matching in RCM

Out of four matching methods in RCM using propensity scores, only two methods will be utilized for this study: *exact* and *caliper* matching. Based on collected CVs between the two groups, propensity scores are calculated using logistic regression (Rosenbaum & Rubin, 1983). In order to calculate the *propensity* scores, the logistic regression model will be used as described below:

Logit(Ti)= Primary diagnosis+ChCom+FIMA+MUtil+StAnA+SF12A+

Age+gender+weight, where T_1=IRF and T_2=SNF – 3.1

The dependent variable of this model is the two comparison groups (IRFs and SNFs), and the independent variables are primary diagnosis and a few of the admission scores (Charlson Comorbidity scores, Functional Independent Measure, Medical Utilization, Anxiety/Depression, and SF-12), as well as demographic variables—age, gender, and weight. The secondary diagnosis (subcategories of primary diagnosis in Table 3.2) is not included to calculate propensity scores because of insufficient degrees of freedom in the logistic regression model for each level of three categorical variables:

primary diagnosis with two levels, gender with two levels, and secondary diagnosis with 12 levels—eight in cardiac diagnosis and four in pulmonary diagnosis.

Since the propensity scores are continuous numbers, it is often impossible to match the exact values from IRFs to SNFs. For example, the propensity score of one participant from SNFs is 0.8513, but there may be no exact matching propensity value found from IRFs. Therefore, the closest value will be matched between them. For this study, exact matching will be called *1:1 matching* since it is not practical to match exact propensity scores between comparison groups with a small number of cases just as in *"hand-matching."* Using *1:1 matching*, 27 participants from SNFs are matched to 27 from IRFs. Ideally, the matched datasets and the results of the analysis using *propensity matching* should be the same as those of "*hand-matching.*" However, it is expected that both the matched datasets and the results of the analysis will be different.

Caliper matching using propensity scores is done in the same manner as caliper matching in the *"hand-matching"* method. One participant from SNFs will be matched with several from IRFs in the ratio of 1:r. For example, with the ratio 1:4, one participant from SNFs will be matched with four participants from IRFs. As a result, 27 patients from SNFs will be matched with 108 out of 319 from IRFs. The difference in propensity scores between the participants from IRFs and SNFs should be less than 2 standard deviation (sd) of the propensity score in order to ensure background characteristics are close. Therefore, the 2 sd inclusion criterion may reduce the number of matched participants from IRFs, and, as a result, the matched dataset using caliper matching may be smaller than the intended ratio of 1:4.

Table 5

Datasets after Matching

MANCOVA	Fisher's Hand Matching	RCM Propensity Matching
Compare results	Data by 1:1 Matching	Data by 1:1 Matching
Compare results	Data by Caliper Matching	Data by Caliper Matching

Using the two matching methods, 4 datasets are prepared. Through *1:1 matching*—27 participants from IRFs are matched to 27 from SNFs—two datasets are produced: one through "*hand-matching*," and another through *propensity matching* just as in Table 5. Through *caliper matching*, two datasets are produced in the same manner, but the number of participants from IRFs will be larger than in *1:1 matching*. After matching, each of the four datasets described in Table 5 will be analyzed separately using the multivariate analysis of covariance (MANCOVA) model.

Data Analysis with Matched Datasets

The analysis will be done using each dataset after a matching method has been applied to the original data. This section describes the primary plan to analyze the matched data based on the hypothesis of the former study, which was not conducted due to the sample size disparity. The hypothesis of the former study was to evaluate whether patients with cardiac and pulmonary diagnoses undergoing rehabilitation interventions administered at IRFs experience better outcomes than patients treated by SNFs. The

results of the analysis will be reported and compared between Fisher's and RCM matching methods.

Rubin and Thomas (2000) suggested the use of analysis of covariance (ANCOVA) or multivariate analysis of covariance (MANCOVA) to adjust the remaining differences in the distribution of covariates between two comparison groups after matching. Intuitively, it may be assumed that pretest scores are associated with posttest scores. ANCOVA or MANCOVA is a way to adjust for any error variance from the association between the pre- and posttest scores by treating pretest scores of all covariates equal in all groups (Hays, 2005). As two separate outcomes, the posttest functional independence measure (FIMD) and the posttest quality of life SF-12D, will be concurrently analyzed to compare between IRFs and SNFs using MANCOVA as described in Table 6.

Table 6

Variables in MANCOVA

Study Components	CVs and IV	DVs
Physical/Functional Assessment	Measure (FIMA)	Functional Independent Measure (FIMD)
Psychological Assessment	Quality of life (SF12A)	Quality of life (SF12D)
Independent Variable	IRFs vs SNFs	

MANCOVA has several advantages over ANCOVA when there are several dependent variables (DVs), especially in protecting from type I error due to multiple tests of correlated DVs (Tabachnick & Fidell, 2007). Again, the pretest scores of FIM and SF12 are used as *covariates* (CVs) in MANCOVA. However, the correlations between

four variables, pre- and posttest scores of FIM and SF-12, will be checked for two reasons: first, whether there are any significant correlations between CVs, and second, whether there are any significant correlations between CVs and the outcomes. In MANCOVA with two groups, the main and interaction effects between two outcomes, FIMD and SF-12D, will be entered into the model.

In MANCOVA, the models are expected to meet three assumptions: normality, linearity, and homogeneity of variance. These assumptions are checked for model-fit and potential transformations may be done to fit the model. Often, however, the assumptions are violated in complex and/or small data. Tabachnick and Fidell (2007) recommend utilizing a logistic regression model in the case of unequal sample sizes between groups and/or assumptions of variance that are not feasible. Logistic regression requires no data assumptions of the models (Agresti, 2007). Therefore, this study proposes to utilize the logistic model as an alternative model if the MACOVA model does not satisfy the parametric assumptions.

The results of the proposed analysis are used to compare two matching methods, Fisher's and RCM, using the same matching approach. For example, a matched dataset using *1:1* matching from *Fisher's* design is compared with another matched dataset using *1:1* matching from *RCM*. As well, a matched dataset using *caliper* matching from *Fisher's* is compared with a dataset using *caliper* matching from *RCM* (Table 5).

CHAPTER FOUR

RESULTS

This study has two goals: first, utilizing two matching methods—Fisher's "hand-matching" and RCM propensity matching—and second, comparing the results of analyses on datasets made by the two matching methods. The two matching methods were brought into this study in order to resolve a problem of the dataset namely, extremely unbalanced numbers of participants between two comparison groups. Due to the discrepancy, the analysis for the main hypothesis in the formal study was not conducted (Beacham, 2008). This study utilized the two matching methods for the dataset with unbalanced participants and hypothesizes that both methods, Fisher's and RCM, would produce similar results in a MANCOVA analysis by using the same matching techniques—1:1 and caliper.

In this chapter, there are fundamentally three sections: 1) preliminary analysis of the original data before matching—not all the data collected for the formal study is used for this present study, 2) datasets created by each matching method, and 3) results of the analysis for each matched dataset. While doing the preliminary analysis, a few issues are brought up related to the reliability of the data before matching. The issues and their resolution are discussed in the first section to follow. After data cleaning based on the

preliminary analysis, matching procedures will be presented, and then each matched dataset will be analyzed according to the analytic plan provided in chapter 3, using the computer program PASW 17.0 (SPSS, INC., Chicago, USA).

The 1:1 matching technique will produce the same or a close number of participants for the comparison facilities, but caliper matching will not. Therefore, datasets made through caliper matching, when analyzed, will weight cells by their sample sizes to adjust for unequal number of participants. The results of the analysis will be written in the order of analysis: first, "hand-matching," and second, propensity matching.

Preliminary Analysis

Table 7 presents descriptive statistics for the two dependent variables and seven covariates (CVs) by type of facility. The two DVs are the discharge (posttest) scores of FIM (FIMD) and SF-12 (SF12D). The seven CVs are the admission (pretest) scores on Charlson Comorbidity Index (ChCom), Functional Independence Measure (FIMA), Medical Utilization (MUtil), Stress/Anxiety (StAnA), SF-12 (SF12A), age, and weight. There are three more CVs, gender, primary and secondary diagnosis, which are not in this table. The primary and secondary diagnoses were described in Table 3 in detail. Therefore, altogether ten CVs will be utilized to match participants between the two facilities. After creating a dataset through each matching method, only two out of ten CVs, FIMA and SF12A, will be used for the analysis in comparing two facilities (IV) on two DVs, FIMD and SF12D, in the MANCOVA model.

Table 7

Description of the Dataset before Matching

		FIMA	SF12A	MUtil	ChCom	StAnA
IRFs	n=318(missing)	316(2)	297(21)	308(10)	306(12)	301(17)
	Mean	***82.88***	344.32	***15.48***	2.474	***9.57***
	SD	12.34	157.48	10.20	1.90	4.01
	Range	74.00	800.00	40.00	10.00	18.00
SNFs	n=27(missing)	24(3)	27(0)	27(0)	26(1)	26(1)
	Mean	79.63	***408.89***	12.93	***2.65***	7.73
	SD	15.59	170.73	10.50	2.04	4.07
	Range	65.00	660.00	35.00	7.00	15.00
		Age	Weight	FIMD	SF12D	
IRFs	n=318(missing)	314(4)	290(28)	311(7)	237(8)	
	Mean	71.98	***180.37***	***105.16***	394.94	
	SD	10.59	56.26	14.87	142.22	
	Range	64	361.40	106.00	660.00	
SNFs	n=27(missing)	27(0)	26(1)	22(5)	18(9)	
	Mean	***76.59***	167.18	98.45	***458.61***	
	SD	12.78	43.40	18.58	160.35	
	Range	52	193.00	66.00	535.00	

Notes: Italic & bold= the larger of the two means, Underlined=DVs

Comparing facilities in Table 7, IRFs have larger means on FIMA, MUtil, StAnA, Weight, and FIMD, while SNFs have larger means on SF12A, ChCom, Age, and SF12D. Comparing pretest scores of FIM (FIMA) and SF-12 (SF-12A), and posttest scores of FIM (FIMD) and SF-12 (SF-12D), both means and SDs increased from pre- to posttest in both IRFs and SNFs except for the SD of SF-12D in SNFs. The means of FIM and SF-12 increased from pre- to posttest within each type of facility: for FIM, 22.28 points in IRFs and 18.82 points in SNFs, and for SF-12, 50.6 points in IRFs and 49.73 in SNFs. In both

IRFs and SNFs, there were more female than male patients (male=143 and female=172 in IRFs, male=9 and female=18 in SNFs).

Most of the variables have less than 5% missing values except SF12A (6.6% missing) and weight (8.8% missing) in IRFs, and FIMD (18.5% missing), and SF12D (33.3% missing) in SNFs. Since there was a smaller number of participants in SNFs but more than enough from IRFs to match between the two facilities, further investigation in SNFs was done in order to deal with missing values.

Table 8

All Participants from SNFs by CVs

Number of cases	FIMA	SF12A	MUtil	ChCom	StAnA	Age	Weight	Gender
1	--	missing	--	--	--	--	--	--
2	missing	--	--	--	--	--	--	--
1	missing	--	--	--	--	--	--	--
1	--	--	--	--	missing	--	--	--
1	--	--	--	missing	--	--	--	--
1	--	--	--	--	--	--	missing	--
20	--	--	--	--	--	--	--	--

Table 8 presents the missing data pattern for SNFs in the entire data on eight CVs. For *matching participants* from SNFs to IRFs, however, those participants with missing values are *not* eliminated. As a result, the total number of participants to be matched is 345: 318 for SNFs, and 27 for IRFs (The final number in here is different from the final number in the original data because this is only a part of the whole data). No missing values were imputed in any CVs or in DVs.

Conventionally, missing values are imputed before matching data by using various imputation methods: maximum likelihood imputation (ML), multiple imputations (MI), mean-median imputation, and last observation carried forward imputation (LOCF). The MatchIt software also requires all missing value fields to be filled before using the program to match data. For this study, it was originally planned to use the MatchIt software to match the data. There are, however, five reasons for not utilizing any imputation procedure for missing values in order to prevent artificial influence to the original data and, ultimately, to the final analysis: First, mean-median imputation and LOCF reduce variance, especially with a small number of participants; second, adding another step of manipulation with parameterized ML or MI would influence the raw data even before matching the data; third, the advantage of Fisher's hand matching method is utilizing the raw data; fourth, using raw data without imputation resembles a "real life" situation; and fifth, any imputation procedure would produce incomparable datasets. Given these reasons, for *hand-matching* in this study, no imputation for missing values will be done because the imputation would dilute the strength of hand-matching that is using only raw values. If any imputation procedure is done to the present dataset, only a new dataset made through propensity matching will be imputed for this present study. As a result, the imputation procedure will be applied to one of the two different matching techniques.

Applying the imputation to only one matching method would produce very different datasets made by the two comparison technique: one dataset has only raw data while another dataset has imputed values added to the raw data. Therefore, the results of

the analysis between the two datasets made by two different matching techniques, hand-matching and propensity matching, cannot be compared. Given these justifications, the datasets will be matched without any imputation for analysis. Further reduction of sample size is expected while matching and analyzing the data due to missing values in several variables.

Matched Datasets

Fundamentally, this study imports the ideas from two matching methods: *Fisher's hand-matching* and *RCM propensity matching*. In Fisher's method, there are two matching techniques: *1:1* and *caliper*. In the RCM method, there are two techniques with identical names to those in Fisher's: *1:1* and *caliper*. As a result, four datasets were created by using two different matching methods: two from Fisher's and two from RCM (See Table 9). Since the names of the two techniques from each matching method are identical, a specific name is given to each dataset after matching.

Table 9

Four Matched Datasets

Methods / Techniques	Fisher's Hand-Matching Method	RCM Propensity Matching Method
1:1 matching	Fisher's 1:1 Matched Data (1)	RCM 1:1 Matched Data (2)
Caliper Matching	Fisher's Caliper Matched Data (3)	RCM Caliper Matched Data (4)

In both matching methods, Fisher's and RCM, participants were matched based on CVs. In the RCM matching method, participants were matched with propensity scores

which are calculated using CVs, and in Fisher's matching method, participants were matched with raw values of CVs. Although the participants from two groups, SNFs and IRFs, might have missing values on either one or both outcomes and/or CVs, they were still utilized and matched. It is because this simulates a "real life" situation with missing values. When an imputation procedure is done in 1:1 matched data, it is assumed that there are equal number of patients in IRFs and SNFs. But in reality, no equal number of patients will be admitted to both facilities.

Fisher's Hand-Matching

As displayed in Table 9 above, two matched datasets are created by using Fisher's hand-matching method: Fisher's 1:1 matched data and caliper matched data. For *Fisher's hand-matching*, a priority among CVs is given. Ho et al. (2007) mentioned that preprocessing matching has a "curse of multi-dimensionality." Due to multiple CVs with many participants in a vector form, it is almost impossible to match many values of CVs at the same time. For example, age and gender might be able to be matched, but weight may not be able to be matched. Therefore, a priority among covariates is given in hand-matching after consulting two experts in the area of medicine and rehabilitation. The priorities are in the following order: primary and secondary diagnosis, pretest scores of FIM and SF-12, Medical Utilization, Charlson Comorbidity Index, Stress/anxiety, and demographics (age, weight, and gender). Based on this order, the data are matched. When there is not a case with exact matching value, the case with the nearest value will be used.

What follows is the *matching procedure* done with the priority of CVs in both 1:1 and caliper matching techniques in Fisher's method. First, participants were divided into two groups based on the *primary* diagnosis: cardiac or pulmonary. Second, each group was divided again into subgroups according to *secondary* diagnosis type: 10 in *cardiac* and four in *pulmonary*. However, in *secondary* diagnoses, the number of patients from IRFs is smaller than that of SNFs for *congestive heart failure* as shown in Table 3.2. Since there are only 9 patients with *congestive heart failure* in IRFs, there are not enough patients to match with all 12 patients with the *same secondary* diagnosis in SNFs. Therefore, the unmatched participants who were diagnosed with the *other secondary* diagnoses from IRFs will be matched to the participant with *congestive heart failure* in SNFs. Third, the rest of the CVs will be matched in the order of the priority mentioned above. Since CVs are continuous variables, the closest value of each CV will be matched between the participants in SNFs and IRFs except for gender. However, when there is a missing value in any CV, the next CV in order of priority will be utilized to match. For example, there are three missing values in the pretest scores of FIM in Table 4.1.2. In that case, the next CV in priority—pretest score of SF-12—will be used to match between SNFs and IRFs.

For both 1:1 and caliper matching techniques, a match should be no more than two SDs of a mean in difference. If no value is found within two SDs in any CVs, no participant from IRFs will be matched to SNFs. As a result, 27 participants from SNFs are matched with 27 from IRFs in 1:1 matching technique, and 27 from SNFs are

matched with 104 from IRFs in caliper matching technique, making 1:3.8 ratio which is less than the planned ratio of 1:4.

RCM Propensity Matching

RCM propensity matching is a simpler way to match because the logistic regression model (Formula 3.1.1) provides a propensity score for each participant. However, the logistic regression model only produced propensity scores for 21 of 27 participants in SNFs. This is because the logistic regression model cannot calculate the probability scores for a participant when any one CV value is missing. As a result, only 21 participants from SNFs will be utilized for both RCM propensity matching techniques, *1:1* and *caliper*.

For RCM propensity *1:1 matching* technique, the criterion of matching two groups of participants is within one-half SD (0.041) of the propensity score. All but one of the 21 participants from SNFs are matched according to this criterion. Although the propensity score of the one participant from SNFs is 0.269 greater than that of the participants from IRFs—over one-half SD—it is retained in order to utilize the maximum possible participants from SNFs. Therefore, in RCM 1:1 matching technique, 21 participants in SNFs were matched to 21 from IRFs with the nearest propensity scores.

For RCM propensity *caliper matching*, the criterion of matching two groups of participants was within one SD of the propensity score (0.082) in order to match as many possible participants from IRFs with the ratio 1:4. However, there were not enough participants in the higher range of the propensity scores (0.2 to 0.5) from IRF to match

participants from the same range of propensity scores from SNFs with the ratio 1:4. For example, there are six participants in the range of 0.20 to 0.43 in SNFs, but there are only 14 participants with the same range of propensity scores in IRFs. As a result, 76 participants from IRFs were matched to 21 from SNFs.

This study hypothesizes that the datasets produced from the same matching technique (*1:1*or *caliper*) in both matching methods (Fisher's and RCM) produce similar results of the same analysis. Before the analysis, the datasets created by the same matching technique were investigated because the matched datasets should contain a large number of the same participants after matching in order to produce similar results of the same analysis. Since participants in SNFs are the same in all four created datasets but only participants in IRFs can be changed in accordance with the two matching methods, the participants in IRFs with the same matching technique are compared. For example, the participants from *Fisher's 1:1 Matched Data* are compared to the participants from *RCM 1:1 Matched Data* (Table 9).

Differences and/or Similarities between Two Matched Datasets

Two matching methods created four datasets (see Table 9). In this section, two comparisons of datasets were done for the same matching technique between the two different matching methods: first, Fisher's *1:1* Matched Data will be compared with RCM *1:1* Matched Data, and second, Fisher's *Caliper* Matched Data will be compared with RCM *Caliper* Matched Data. Descriptive statistics for the four datasets will be

reported within two separate tables based on *matching technique*, including n, means, SD, and range.

Table 10

Description of the Datasets Created through 1:1 Matching Technique

Fisher's/RCM		FIMA	SF12A	MUtil	ChCom	StAnA
Fisher's 1:1	n=27(missing)	27	27	27	26(1)	27
	Mean	***82.59***	377.69	***15.78***	2.19	***9.89***
	SD	13.32	141.07	10.66	1.65	3.97
	Range	54.00	522.50	36.00	6.00	13.00
RCM 1:1	n=27(missing)	21	21	21	21	21
	Mean	78.80	***414.76***	13.76	***2.43***	8.29
	SD	15.26	161.59	12.14	2.25	5.04
	Range	48.00	510.00	39.00	10.00	16.00
Fisher's/RCM		Age	Weight	FIMD	SF12D	
Fisher's 1:1	n=27(missing)	27	26 (1)	27	27	
	Mean	71.48	181.93	***<u>108.259</u>***	<u>415.7143</u>	
	SD	12.72	47.23	9.928	138.23	
	Range	42	202.20	39.00	497.50	
RCM 1:1	n=27(missing)	21	21	21	17	
	Mean	***72.67***	***182.00***	<u>103.95</u>	***<u>423.38</u>***	
	SD	10.46	55.98	13.28	126.51	
	Range	34	218.50	48.00	397.50	

Notes: Italic & bold=the larger of the two means, Underlined=DVs

Table 10 describes the two datasets created through *1:1 matching technique* in all CVs except primary and secondary diagnosis. Fisher's 1:1 Matched Data has greater means in FIMA, MUtil, StAnA, and FIMD than RCM 1:1 Matched Data, while RCM 1:1 Matched Data has greater means in SF12A, ChCom, age, weight, and SF12D.

However, the means of all variables are within one SD of each other, which clarifies that there is no significant difference between the two datasets in the descriptive statistics. Unexpectedly, when 27 participants in Fisher's 1:1 Matched Data were compared with 21 participants in RCM 1:1 Matched Data, only two participants were in both datasets.

Table 11

Description of Datasets Created through Caliper Matching Technique

		FIMA	SF12A	MUtil	ChCom	StAnA
Fisher's Caliper	n=104(missing)	104	104	104	97(7)	102(2)
	Mean	***82.04***	391.42	14.08	***2.03***	***9.18***
	SD	12.57	148.69	10.43	1.839	4.206
	Range	59.00	750.00	39.00	10.00	18.00
RCM Caliper	n=76(missing)	76	76	76	76	76
	Mean	81.60	***406.12***	***14.27***	2.00	8.48
	SD	13.99	164.78	10.51	1.904	3.95
	Range	70.00	712.50	39.00	10.00	16.00
		Age	Weight	FIMD	SF12D	
Fisher's Caliper	n=104 (missing)	104	99(5)	102(2)	79(25)	
	Mean	***73.82***	***182.239***	***105.558***	418.89	
	SD	10.47	51.78	15.88	135.03	
	Range	54	331.00	105.00	600.00	
RCM Caliper	n=76(missing)	76	76	74(2)	56(20)	
	Mean	73.91	173.34	105.33	***439.55***	
	SD	10.55	49.49	16.09	144.16	
	Range	49	230.90	106.00	597.50	

Note: Italic & bold=the larger of the two means, Underlined=DVs

Table 11 describes the two datasets created through *caliper matching technique* in all CVs except primary and secondary diagnosis. Fisher's Caliper Matched Data has greater means in FIMA, ChCom, StAnA, age, weight, and FIMD than RCM Caliper

Matched Data, while RCM Caliper Matched Data has greater means in SF12A, MUtil, and SF12D. However, the means of all variables are within one SD of each other, which clarifies that there is no significant difference between the two datasets in the descriptive statistics. Unexpectedly, when 104 participants in Fisher's Caliper Matched Data were compared with 76 participants in RCM Caliper Matched Data, only 32 participants were in both datasets.

Even though there are mostly different participants in the two datasets, notice that there are similar descriptive statistics when using the same matching technique between Fisher's and RCM. Since there are mostly different participants in IRFs in each dataset, one might expect that the results of the analyses of the two datasets in the same matching technique to be different. However, the descriptive statistics showed that they are similar. The analyses of datasets will provide whether the differences are statistically significant or not using MANOCVA. Each dataset will be analyzed, testing the differences between two facilities (IV) on the two DVs—posttest scores of FIM and SF-12—after adjusting for two CVs: pretest scores of FIM and SF-12.

Results of the Analysis

After matching and before the analysis of each dataset, descriptive statistics of the two facility types will be given, and the correlations between the two DVs and the two CVs will be investigated. The two DVs are posttest scores of FIM (FIMD) and SF-12 (SF12D), and the two CVs are pretest scores of FIM (FIMA) and SF-12 (SF12A). The sample size (n), means, SD, and range of each variable will be in the descriptive

statistics. It is expected that the pretest scores have a large correlation with the posttest scores and that there is no strong correlation between CVs. Tabachnick and Fidell (2007) recommended, “Any CV with a squared multiple correlation (SMC) in excess of 0.50 may be considered redundant and deleted from further analysis” (p. 201). Therefore, if the SMC is greater than 0.50, one of the two CVs will be eliminated from the analysis in the order of the theoretical importance which is recommended by the two experts in health and rehabilitation.

After evaluation of the correlations among the two DVs and two CVs, the final model will be established with the CV(s) that should be in the model. Then, each dataset will be analyzed using the MANCOVA model in Table 12. The result of each analysis will be reported in this chapter.

Table 12

MANCOVA Model

MANCOVA Model	**Dependent Variables (DVs)**	**Independent Variable (IV)**	**Covariates (CVs)**
Variables	Posttest FIM Posttest SF12	Facility (1=IRF, 0=SNF)	Pretest FIM Pretest SF12

For each MANCOVA model, the tenability of the assumptions for normality of sampling distributions, homogeneity of variance-covariance, and linearity were evaluated. Normality (degrees of freedom (*df*) is greater 20 for error sum of squares) and linearity of MANCOVA will be evaluated according to the recommendations of Tabachnick and Fidell (2007). Homogeneity of variance-covariance will be evaluated

through Box-M. As well, Wilks' Lambda will be reported for the analysis of the data made through 1:1 matching technique from both Fisher's and RCM methods for the two groups of participants in both facilities are presumably equal or at least close to each other, while Pillai's Trace will be reported for the analysis of the datasets made through caliper matching technique. Pillai's Trace is used because it is a better estimate for non-experimental data with unequal sample sizes (Olson, 1979). When significant differences are found between facilities on the two DVs in the MANCOVA, further investigations will be done by looking at two ANCOVA models with separated DVs with the same CVs and IV. The results of the analyses will be reported in the following order of datasets: Fisher's 1:1 Matched Data, RCM 1:1 Matched Data, Fisher's Caliper Matched Data, and RCM Caliper Matched Data.

Results of Fisher's 1:1 Matched Data

Table 13 and Table 14 provide the descriptive statistics comparing between IRFs and SNFs. Due to missing values, the number of participants is reduced, especially in the posttest score of SF-12: from 27 to 21 in IRFs and from 27 to 18 in SNFs. Sample sizes in IRFs are bigger than that of SNFs in three out of four variables—FIMA, FIMD, and SF12A—because there are more missing values in SNFs. Therefore, the final sample size for the MANCOVA analysis became 18. These differences; however, may reflect what happens in the real world: more patients may be admitted to one facility than another; or discharged earlier than expected disrupting collection of prospective research data.

Comparing the two facilities, the participants in IRFs have larger means in both pre- and posttest scores of FIM than those in SNFs, while the participants in SNFs have larger means in both pre- and posttest scores of SF-12 than those in IRFs (Table 13 and Table 14). This may mean that there is a relationship between pre- and posttest scores in both FIM and SF-12: When the pretest score was low, the posttest score was low, and when the pretest score was high, the posttest score was high. In addition, both FIM and SF-12 scores increased between pre- and posttest in both facility types: the mean difference was 25.66 points for FIM and 38.03 for SF-12 in IRFs, and 18.83 points for FIM and 49.73 for SF-12 in SNFs.

Table 13

Descriptive Statistics of Fisher's 1:1 Matched Data for IRFs

	FIMA	SF12A	FIMD	SF12D
N (missing)	27	27	27	21(6)
Mean	***82.59***	377.68	***108.25***	415.71
Std. Deviation	13.32	141.07	9.928	138.26
Range	54.00	522.50	39.00	497.50

Table 14

Descriptive Statistics of Fisher's 1:1 Matched Data for SNFs

	FIMA	SF12A	FIMD	SF12D
N (missing)	24	27	22(5)	18(9)
Mean	79.62	***408.88***	98.45	***458.61***
Std. Deviation	15.59212	170.72573	18.58245	160.34710
Range	65.00	660.00	66.00	535.00

This shows that both facilities improved functional independence and quality of life of the patients through their "usual care" of rehabilitation. However, IRFs increased in FIM scores more than SNFs, while SNFs increased in SF-12 scores more than IRFs. Since each facility does better on one of the two DVs than the other, it may be the case that both facility types were equally effective in rehabilitating patients.

Table 15

Correlations among DVs and CVs of Fisher's 1:1 Matched Data for IRFs

	FIMA	SF12A	FIMD	SF12D
FIMA	1	.412*	.571**	.109
SF12A	.412*	1	.092	.687**
FIMD	.571**	.092	1	.138
**p<0.01, *p<0.05 (2-tailed).				

Table 16

Correlations among DVs and CVs of Fisher's 1:1 Matched Data for SNFs

	FIMA	SF12A	FIMD	SF12D
FIMA	1	.195	.803**	-.017
SF12A	.195*	1	.254	.780**
FIMD	.803**	.254	1	.257
SF12D	-.017	.780**	.257	1
**p<0.01, *p<0.05 (2-tailed).				

Table 15 and Table 16 describe the correlations among CVs and DVs in IRFs and SNFs respectively. A significant correlation is found between two CVs: FIMA and SF12A, (r =0.412, p<0.05) only in IRFs. Although the correlation is statistically significant, no CV is taken out from the MANCOVA model to remedy the multicollinearity problem among CVs because the SMC is less than 0.50 (SMC=0.170) according to the plan recommendations by Tabachnick and Tidell (2007). The significant correlations between pre- and posttest scores of FIM (r =0.571 for IRFs, r =0.803 for SNFs, both p<0.01) and SF-12 (r =0.550 for IRFs, r =0.780 for SNFs, both p<0.01), show that CVs are adequately reliable for the MANCOVA model.

Table 17

Results of MANCOVA Model Fisher's 1:1 Matched Data

Effect	Wilks' Lambda	F-value	Test df	Error df	p-value
Intercept	0.512	15.724	2	33	<0.0001
FIMA	0.545	13.753	2	33	<0.0001
SF-12A	0.577	12.120	2	33	<0.0001
Facility	0.860	2.679	2	33	0.082

Note: This is where author provide extra information important to the data, such as findings that approach statistical significance depending on the p value: Significant at the p<0.05 level.

Although the number of participants (n) is only 18 out of 27 for SNFs in SF-12D, the model assures multivariate normality with more than 20 *df* for error sum of squares based on Table 17. There are no outliers in both groups because the ranges are within three SDs of the mean, which satisfies the linearity and variance assumptions. Box-M is

tested for the homogeneity assumption, which confirms the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups ($F(3, 402399)=2.073$, $p=0.12$). With the use of Wilks' criterion, there is no significant difference between the two facilities on the two combined outcomes (DVs) after adjusting for the two CVs.

The partial η^2 of IV (Facility) is 0.14 (1-Wilks' Lambda), which explains that the association with DVs is small, but a larger association was found between DVs and both CVs: partial $\eta^2 = 0.455$ (1-0.545) for FIMA, and partial $\eta^2 = 0.423$ (1-0.577) for SF12A. The results suggests that there is no difference between IRFs and SNFs in rehabilitating this group of patients with cardiac and/or pulmonary diagnosis after controlling for the pretest scores of FIM and SF-12 ($F(2,33)=2.679$, $p=0.082$) (Table 17).

Results of RCM 1:1 Matched Data

Table 18 and Table 19 describe the data between IRFs and SNFs. Due to the missing propensity scores in SNFs, only 21 participants in SNFs are matched to 21 in IRFs. The number of participants is reduced further—17 out of 21 in IRFs and 18 out of 21 in SNFs—in the posttest score of SF-12. Sample sizes in SNFs are bigger than that of IRFs in all four variables because no participants are eliminated even though there are missing values or missing propensity scores from SNFs.

Comparing the two facilities, the participants in IRFs have larger means in SF-12A and FIMD than those in SNFs, while the participants in SNFs have larger means in FIMA and SF-12D than those in IRFs (Table 18 and Table 19). In addition, both FIM

and SF-12 scores increase from pre- to posttest in both facility types: mean difference was 25.15 for FIM and 8.62 for SF-12 in IRFs, and 18.83 for FIM and 49.73 for SF-12 in SNFs. This shows that both facilities improved functional independence and quality of life of the patients through their "usual care" of rehabilitation. However, IRFs increased in SF-12 scores more than SNFs, while SNFs increased in FIM scores more than IRFs. Since each facility does better on one of the two DVs than the other, it may be the case that they are equally effective in rehabilitating patients.

Table 18

Descriptive Statistics of RCM 1:1 Matched Data for IRFs

	FIMA	SF12A	FIMD	SF12D
N (missing)	21	21	21	17 (4)
Mean	78.80	***414.76***	***103.95***	423.38
Std. Deviation	15.25	161.58	13.27	126.51
Range	48.00	510.00	48.00	397.50

Table 19

Descriptive Statistics of RCM 1:1 Matched Data for SNFs

	FIMA	SF12A	FIMD	SF12D
N (missing)	24	27	22 (2)	18
Mean	***79.62***	408.88	98.45	***458.61***
Std. Deviation	15.59	170.72	18.58	160.34
Range	65.00	660.00	66.00	535.00

Table 17 describes the correlations among CVs and DVs in IRFs only because the correlations in SNFs are the same as Table 4.3.3b. A significant correlation is found between two CVs, FIMA and SF12A, ($r = 0.676$, $p < 0.005$). Although the correlation is

statistically significant, no CV is taken out from the MANCOVA model to remedy the multicollinearity problem among CVs because the SMC is less than 0.5 (SMC=0.457) according to the plan recommended by Tabachnik and Tidell (2007). The significant correlation between pre- and posttest scores of FIM (r =0.639, p<0.01) and SF-12 (r =0.630, p<0.01), shows that CVs are adequately reliable for the MANCOVA analysis in Table 20.

Table 20

Correlations of DVs and CVs of RCM 1:1 Matched Data

	FIMA	SF12A	FIMD	SF12D
FIMA	1	.676**	.639**	.397
SF12A	.676**	1	.129	.630**
FIMD	.639**	.190	1	.194
SF12D	.397	.630**	.194	1
**p<0.01, *p<0.05 (2-tailed).				

Although n is only 17 out of 21 for IRFs in SF-12D, the model assumes multivariate normality with more than 20 df for error sum of squares based on Table 18. There are no outliers in both groups because the ranges are within 3 SD of the mean, which satisfies the linearity assumption. Box-M is tested for the homogeneity assumption, which confirms the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups (F(3, 184320)=0.624, p=0.60). With the use of Wilks' criterion, there is no significant difference between two facilities—IRFs

and SNFs—on the two DVs, FIMD and SF12D, after adjusting for the two CVs, FIMA and SF12A (Table 21).

Table 21

Results of MANCOVA Model in RCM 1:1 Matched Data

Effect	Wilks' Lambda	F-value	Test df	Error df	p-value
Intercept	0.496	14.758	2	29	<0.0001
FIM A	0.464	16.752	2	29	<0.0001
SF-12A	0.508	14.025	2	29	<0.0001
Facility*	0.912	1.402	2	29	0.262

Note: This is where author provide extra information important to the data, such as findings that approach statistical significance depending on the p value: Significant at the $p<0.05$ level.

The partial η^2 of IV (Facility) is very small, 0.088 (1-Wilks' Lambda), which implies a small association with DVs, but a larger association was found between DVs (the combined outcomes) and both CVs: partial η^2 =0.536 (1-0.464) for FIMA, and partial η^2 =0.492 (1-0.508) for SF12A. The results are the same as in Fisher's 1:1 Matched Data: there is no difference between IRFs and SNFs in rehabilitating this group of patients with cardiac and/or pulmonary diagnosis after controlling for the pretest scores of FIM and SF-12.

Results of Fisher's Caliper Matched Data

Table 22 and Table 23 describe the data between IRFs and SNFs. As described in the data matching section, the number of participants is reduced due to missing values, especially in the posttest score of SF-12: 18 out of 27 in SNFs and 56 out of 104 IRFs,

making the ratio from 1:4 to 1:3.1. All 27 participants in SNFs are utilized to match to 104 participants from IRFs, but there are missing values in both SNFs and IRFs.

Table 22

Descriptive Statistics of Fisher's Caliper Matched Data for IRFs

	FIMA	SF12A	FIMD	SF12D
N (missing)	104	104	102 (2)	79 (25)
Mean	***82.59***	391.41	***105.55***	418.89
Std. Deviation	12.56	148.68	15.88	135.03
Range	54.00	750.00	105.00	600.00

Table 23

Description of Fisher's Caliper Matched Data for SNFs

	FIMA	SF12A	FIMD	SF12D
N (missing)	24	27	22	18
Mean	79.62	***408.88***	98.45	***458.61***
Std. Deviation	15.59	170.72	18.58	160.34
Range	65.00	660.00	66.00	535.00

Comparing the two facilities, the participants in IRFs have larger means in both pre- and posttest scores of FIM than those in SNFs, while the participants in SNFs have larger means in both pre- and posttest scores of SF-12 than those in IRFs (Table 22 and Table 23). In addition, both FIM and SF-12 scores increased from pre- to posttest in both facility types: mean difference was 22.96 for FIM and 27.48 for SF-12 in IRFs, and 18.83 for FIM and 49.73 for SF-12 in SNFs. This shows that both facilities improved functional independence and quality of life of the patients through their "usual care" of rehabilitation. However, IRFs increased in FIM scores more than SNFs, while SNFs increased in SF-12 score more than IRFs.

Table 24

Correlations among DVs and CVs of Fisher's Caliper Matched Data

	FIMA	SF12A	FIMD	SF12D
FIMA	1	.279**	.489**	.169
SF12A	.279**	1	.156	.547**
FIMD	.555**	.157	1	.163
SF12D	.169	.547**	.163	1
**p<0.01, *p<0.05 (2-tailed).				

Table 24 describes the correlations among CVs and DVs only in IRFs because the correlations in SNFs are the same as Tables 13b. A significant correlation is found between two CVs, FIMA and SF12A, (r =0.279, p<0.01). Although the correlation is statistically significant, no CV is taken out from the model to remedy for the multicollinearity problem among CVs because the SMC is less than 0.5 (SMC=0.078). The significant correlation between pre- and posttest scores of FIM (r =0.489, p<0.01) and SF12 (r =0.547, p<0.01) shows that CVs are adequately reliable for the MANCOVA model.

Though n is only 18 out of 27 for SNFs in SF-12D, the model assures multivariate normality with more than 20 *df* for error sum of squares based on the Table 25. There are no outliers in both groups because the ranges are within 3 SD of the mean, which satisfies the linearity and variance assumption. Box-M is tested for the homogeneity assumption,

which rejects the null hypothesis that the observed covariance matrices of the dependent variables are not equal across groups (F(3, 11077)=3.102, *p*=0.026).

Table 25

Results of MANCOVA Model in Fisher's Caliper Matched Data

Effect	Pillai's Trace	F-value	Test df	Error df	p-value
Intercept	0.509	46.175	2	89	<0.0001
FIMA	0.333	22.169	2	89	<0.0001
SF-12A	0.329	21.826	2	89	<0.0001
Facility	0.083	4.032	2	89	0.021

The homogeneity assumption is often not feasible with unequal n between comparing groups. However, Levene's test accepts that the null hypothesis that the error variance of the two DVs are equal across groups: F(1, 92)=0.679, *p*=0.42 for SF12D and F(1, 92077)=1.927, *p*=0.168) for FIMD. With the use of Pillai's Trace, there is significant difference between two facilities on the two combined outcomes after adjusting for the two CVs. The partial η_2 of IV (Facility) is very small, 0.083, which implies a small association with DVs, but a larger association was found between DVs and both CVs: partial η^2 =0.333 for FIMA, and partial η^2 =0.329 for SF12A. There is difference between IRFs and SNFs in rehabilitating this group of patients with cardiac and/or pulmonary diagnosis after controlling for the pretest scores of FIM and SF-12 (F(2, 89)=4.032, p=0.21). In order to investigate the impact of the main effect (the difference between

facilities-IV) on the individual DVs, a univariate ANCOVA model for each DV will be utilized.

Table 26

Two ANCOVA Models of Fisher's Caliper Matched Data

Source	Dependent Variable	Type III Sum of Squares	Df	Mean Square
Intercept	SF12D	94913.795	1	94913.795***
	FIMD	7335.423	1	7335.423***
FIMA	SF12D	512.747	1	512.747
	FIMD	3624.691	1	3624.691***
SF12A	SF12D	557084.130	1	557084.130***
	FIMD	83.188	1	83.188
Facility	SF12D	6557.540	1	6557.540
	FIMD	581.420	1	581.420**
Error	SF12D	1138598.255	90	12651.092
	FIMD	7314.787	90	81.275
Total	SF12D	1.882E7	94	
	FIMD	1095390.000	94	
a. R Squared = .357 (Adjusted R Squared = .335) ***p<0.001, ** p<0.01, *p<0.05				
b. R Squared = .385 (Adjusted R Squared = .365)				

With the posttest of SF-12 as the DV in the ANCOVA model, no significant differences were found between the two facilities, while significant differences were found on the main effect with the posttest of FIM as the DV (Table 26). This clarifies that the significant main effect comes from the differences in the posttest scores of FIM between the two facilities. In conclusion, IRFs take better care of the patients with

cardiac and/or pulmonary diagnosis than SNFs as measured by functional independence after controlling for the pretest of FIM and SF-12.

Results of RCM Caliper Matched Data

Table 27

Descriptive Statistics of RCM Caliper Matched Data for IRFs

	FIMA	SF12A	FIMD	SF12D
N (missing)	76	76	74	56
Mean	***81.60***	406.11	***105.33***	439.55
Std. Deviation	13.99	164.78	16.09	144.16
Range	70.00	712.50	106.00	597.50

Table 28

Description of RCM Caliper Matched Data SNFs

	FIMA	SF12A	FIMD	SF12D
N (missing)	24 (3)	27	22 (5)	18 (9)
Mean	79.62	***408.88***	98.45	***458.61***
Std. Deviation	15.59	170.72	18.58	160.34
Range	65.00	660.00	66.00	535.00

Table 27 and Table 28 describe the data between IRFs and SNFs. Due to missing values, the number of participants is reduced, especially in the posttest score of SF-12: 18 out of 21 in SNFs and 56 out of 76 IRFs and, making the ratio from 1:4 to 1:3.1 again. All 21 participants in SNFs are utilized to match to 76 participants from IRFs, but there are missing values in both SNFs and IRFs.

Comparing the two facilities, the participants in IRFs have larger means in both pre- and posttest scores of FIM than those in SNFs, while the participants in SNFs have

larger means in both pre- and posttest scores of SF-12 than those in IRFs in Table 23a and Table 23b. In addition, both FIM and SF-12 scores increased from pre- to posttest in both facility types: mean difference was 23.73 for FIM and 33.44 for SF-12 in IRFs, and 18.83 for FIM and 49.73 for SF-12 in SNFs. This means that there is an increase in both pre- and posttest scores in both FIM and SF-12. This shows that both facilities improved functional independence and quality of life of the patients through their "usual care" of rehabilitation. However, IRFs increased in FIM scores more than SNFs, while SNFs increased in SF-12 scores more than IRFs.

Table 29

Correlations among DVs and CVs of RCM Caliper Matched Data

	FIMA	SF12A	FIMD	SF12D
FIMA	1	.343**	.652**	.114
SF12A	.343**	1	.145	.623**
FIMD	.652**	.145	1	.242
SF12D	. 114	.623**	.242	1
**p<0.01, *p<0.05 (2-tailed).				

Table 29 describes the correlation among CVs and DVs for IRFs only. A significant correlation is found between two CVs, FIMA and SF12A, (r =0.343, p<0.01). Although the correlation is statistically significant, no CV is taken out from the model for the multicollinearity among CVs because the SMC is less than 0.5 (SMC=0.118) according to the plan recommended by Tabachnick and Tidell (2007). The significant correlation between pre- and posttest scores of FIM (r =0.652, p<0.01) and SF-12 (r =0.623, p<0.01) shows that CVs are adequately reliable for the MANCOVA analysis.

Table 30

Results of MANCOVA Mode in RCM Caliper Matched Data

Effect	Pillai's Trace	F-value	Test df	Error df	p-value
Intercept	0.549	46.175	2	67	<0.0001
FIMA	0.508	22.169	2	67	<0.0001
SF-12A	0.449	21.826	2	67	<0.0001
Facility	0.123	4.032	2	67	0.021

Though the number of participants (*n*) is only 18 for SNFs in SF-12D, the model assures multivariate normality with more than 20 *df* for error sum of squares based on the Table 30. There are no outliers in both groups because the ranges are within 3 SD of the mean, which satisfies the linearity and variance assumption. Box-M is tested for the homogeneity assumption, which confirms the null hypothesis that the observed covariance matrices of the dependent variables are not equal across groups (F(3, 13372)=1.753, *p*=0.154). Levene's test also confirms that the null hypothesis that the error variance of the two DVs are equal across groups: F(1, 70)=0.705, *p*=0.404 for SF12 and F(1, 70)=1.116, *p*=0.294). With the use of Pillai's Trace, there is a significant difference between IRFs and SNFs on the DVs after adjusting for the two CVs (F=4.697, p=0.012). The partial η^2 of IV (Facility) is very small, 0.123, which implies small association with DVs, but a larger association was found between DVs and both CVs: partial η^2 =0.508 for FIMA, and partial η^2 =0.449 for SF12A. There is difference between IRFs and SNFs in rehabilitating this group of patients with cardiac and/or pulmonary diagnosis after controlling for the pretest scores of FIM and SF-12 (F(2,

67)=4.032, p=0.021). In order to investigate the impact of the main effect (the difference between facilities-IV) on the individual DVs, a univariate ANCOVA model for each DV will be utilized.

Table 31

Two ANCOVA Models of RCM Caliper Matched Data

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square
Intercept	SF12D	232051.421	1	232051.421***
	FIMD	6348.129	1	6348.129***
FIMA	SF12D	33025.882	1	33025.882
	FIMD	4280.710	1	4280.710***
SF12A	SF12D	631969.680	1	631969.680***
	FIMD	26.319	1	26.319
Facility	SF12D	1252.921	1	1252.921
	FIMD	638.606	1	638.606***
Error	SF12D	817827.678	68	12026.878
	FIMD	5429.031	68	79.839
Total	SF12D	1.575E7	72	
	FIMD	831211.000	72	
a. R Squared = .446 (Adjusted R Squared = .422) ***=p<0.001, ** : p<0.01, *: p<0.05				
b. R Squared = .504 (Adjusted R Squared = .482)				

With the posttest of SF-12 as the DV in ANCOVA model, no significant differences were found between the two facilities, while significant differences were found on the main effect with the posttest of FIM as the DV (Table 31). This clarifies that the significant main effect comes from the differences in the posttest scores of FIM between the two facilities. In conclusion, IRFs are able to gain higher gains in the

patients with cardiac and/or pulmonary diagnosis than SNFs as measured by functional independence, which is the same conclusion of the analysis of Fisher's Caliper Matched Data.

Summary of the Results

In this section, the results are summarized beginning with the descriptive statistics and ending with the four MANCOVA models and two additional ANCOVA models. Fundamentally, the results are to answer the hypothesis of the formal study, which was not conducted due to the discrepancy of the number of participants between the two comparison facilities. The hypothesis of the formal study was comparing the outcomes between patients with cardiac and pulmonary diagnoses undergoing rehabilitation interventions administered at IRFs and SNFs. After this summary, the results of the analysis will be compared and discussed in the next chapter according to the research questions of this present study.

The outcome variables (DVs) were posttest scores of FIM and SF-12, and pretest scores of the same variables were used to adjust for the differences of the participants at baseline after matching. There are four main aspects of the present study that explain the differences between IRFs and SNFs in rehabilitation of the patients. First, two contradicting results were generated from two different matching techniques: Both Fisher's and RCM 1:1 Matched Datasets provided no significant differences between IRFs and SNFs (the main effect) on the two DVs, while both Fisher's and RCM Caliper Matched Datasets showed significant differences. The present study does not attempt to

evaluate the optimal matching method. The contradicting results limit making a definitive conclusion in answering the formal research hypothesis. According to the matching plan, the caliper matching technique enables a researcher to add more participants from IRFs using a ratio of SNFs-to-IRFs. Second, given the significant main effect on the MANCOVA models for both Caliper Matched Datasets, further investigation of two ANCOVA models clarified that IRFs have better effect on patients' functional independence than SNFs. Third, based on the descriptive statistics, both facilities, IRFs and SNFs, increased patients' functional independence and quality of life. Lastly, both measures, FIM and SF-12, positively and statistically correlated between pre- and posttest scores. Moreover, pretest scores of both FIM and SF-12 are significant predictors to posttest scores in all models of MANCOVA and ANCOVA. This means that patients who were admitted with a low or high score were discharged with a low or high score, respectively, regardless of the facility.

It is not clear to find some consistency between datasets made by the same matching technique. Recall, when the participants of the datasets were investigated, they were different in IRFs (the participants from SNFs stay the same throughout). Out of 27 participants in Fisher's 1:1 Matched Dataset and out of 21 participants in RCM 1:1 Matched Dataset, only two participants were the same. Similarly, out of 105 participants in Fisher's Caliper Matched Dataset and out of 76 participants in RCM Caliper Matched Dataset, only 32 participants were the same. One can ask the question: How could this be happening—similar results of the same analysis with different sets of participants? This

question will be investigated in the next chapter according to the research questions of this present study. .

CHAPTER FIVE

DISCUSSION

The goal of this chapter is to discuss the hypothesis and research questions with the results from this present study. The study hypothesized that the two matching methods—Fisher's and RCM—produce similar results with the same analysis with three following research questions: 1) what are the differences, if any, in the results after analysis of the matched datasets produced by the two matching methods, Fisher's and RCM?; 2) what are the similarities, if any, in the results after analysis of the matched datasets produced by the two matching methods, Fisher's and RCM?; 3) by what adjustments could be made in the matching methods to control for the differences found in the results of the same analysis, if possible?

Through analyzing the results of the same analysis in chapter four, three research questions can be parsimoniously boiled down to one question: how could the matched datasets using the same matching technique between the two different matching methods, Fisher's and RCM, produce similar results with different sets of subjects? The question is divided into two points: first, how could the same matching technique between two different methods produce different sets of subjects, and second, how could the datasets with different sets of subjects produce similar results of the same analysis? In addition,

the reasons why there are different results of datasets made by two matching techniques (1:1 and caliper) will be discussed. After the discussion of results, conclusion, implications, and limitations of this study, the recommendations for prospective study topics will be addressed.

How Different Datasets Made With the Same Matching Technique

This present study hypothesized that matched datasets using the two proposed matching methods should produce similar results with the same analysis. It was presumed that the matched datasets using the same matching technique with the two proposed matching methods, Fisher's and RCM, should have similar sets of subjects in order to produce similar results. This was not the case. Only two participants belonged to both datasets made by 1:1 matching technique, and 32 participants belonged to both datasets made by caliper matching technique. How did this happen?

There is one fundamental difference in the same matching technique in the two methods, Fisher's and RCM. The difference in sets of subjects was caused by alternate ways of using CVs between the two matching methods. Fisher's matching method prioritized each CV in matching, while RCM matching method treats all CVs with equal weight in calculating propensity scores. Using Fisher's method, the participants between the two facilities were matched closer with higher-priority CVs than with lower-priority CVs. But, in the datasets made by RCM matching method, the participants between the two facilities are matched based on the propensity score of each person. The propensity scores are estimated using each CV with equal weight. The participants were matched

using the closest propensity score possible between the two facilities. As a result, the two matching methods—one with the priority among CVs and another without—produced different sets of subjects between the two datasets with the same matching technique.

Therefore, there are different sets of subjects with the same matching technique between two methods because Fisher's method had priority among CVs in matching while RCM method did not. It would be possible to produce similar sets of subjects with the same matching technique between the two methods if both methods prioritized CVs or treated CVs equally. As a consequence, the results of the same analysis would be similar.

How Similar Results of Analyses With the Different Datasets

Usually, with different sets of subjects, it is expected to see different results of the same analysis. After matching the data with the same technique in the two methods, however, the results of the datasets were similar with different sets of subjects in this present study. Two possible reasons are discussed in this section.

First, there were similar background characteristics and outcomes between datasets made by the same matching technique. According to Table 10 and Table 11, the descriptive statistics presented were similar between two datasets for seven CVs and two DVs. As a consequence, the results of the MANCOVA model were similar.

Second, the original subjects and the matched datasets have similar background characteristics and outcomes. When Table 7 is compared with Table 10 and with Table

11, similarities are found between them for seven CVs and two DVs. This shows that the participants were chosen to be matched not just from a narrow segment of the original subjects but from its entire range in both matching methods.

In conclusion, the results of the datasets made by Fisher's hand-matching method are the same as that of RCM propensity matching method. This implies that Fisher's hand-matching method is as effective as RCM propensity matching method.

Different Results Between 1:1and Caliper Matched Data

In both Fisher's and RCM method, there are two matching techniques; 1:1 and caliper. 1:1 matching technique matches participants one on one, while caliper matching technique matches with the planned ratio of 1:4 between SNFs and IRFs. After the caliper matching technique was executed, however, the final ratio went down to 1:3.1 in both Fisher's and RCM method.

In the datasets made by Fisher's method, when the descriptive statistics were compared between 1:1 Matched Data (See Table 10) and Caliper Matched Data (See Table 11), the means and SDs were very close to each other. One might expect to see similar results of the same analysis, but the main effect (IV) was different: there was no significant difference between SNFs and IRFs in 1:1 Matched Data, but there was significant difference in Caliper Matched Data. This was mainly because of the increase in the number of participants in IRFs, which caused the reduction of standard error (SE) in the Caliper Matched Data made by Fisher's method; from 25.01 to 12.86 in SF12D, and from 2.20 to 1.03 in FIMD. In the same manner, in the Caliper Matched Data made

by RCM method, SE decreased from 25.26 to 14.82 in SF12D, and from 2.52 to 1.21 in FIMD. The smaller SE makes 95% confidence intervals smaller. As a result, significant difference between IRFs and SNFs was found only in Caliper Matched Data.

This implies, in such a small dataset as this present study, that the caliper matching technique would produce a better estimate by increasing the sample size. According to the Central Limit Theorem, the approximation of a statistical model function improves with a normal distribution in larger sample sizes (Hays, 1994). Lind, Marchal, and Wathen (2008) observed that the distribution of a sample started to become normal when 20 participants were sampled from a subjects regardless of the subjects's distribution, and recommended a sample size of 30 or more. Therefore, 18 participants in SNFs is too small of a number to adequately compare the main effect of taking care of the patients with cardiac and pulmonary diagnosis between the two facilities, IRFs and SNFs. It is clear that the present data has a limitation due to missing values which reduced the total number of participants from 27 to 18 in SNFs. However, the question remains: what is the appropriate way of dealing with the missing values and how?

In chapter four, various reasons were discussed for why any imputation procedure would not be utilized in this study. There were basically three reasons why no imputation procedure was implemented before matching data. First, the advantage of Fisher's matching method was using raw scores. Second, it would produce incomparable datasets between two matching methods, Fisher's and RCM, because the imputation procedure would be applied to only RCM method in order to keep the advantage of Fisher's matching method. As a result, the comparison between the two matching methods is of

no value by using two incomparable datasets; one with imputation and another without. Third, it simulated “real life” between IRFs and SNFs by not imputing missing values. If any one of the imputation procedures was utilized in 1:1 matching technique, it would be assumed that the number of participants was the same between IRFs and SNFs, 27:27, which would not be a “real” but an “ideal” situation. In “real life,” however, the number of participants would be different between IRFs and SNFs for many reasons. For example, a health-care system may limit such patients to be admitted to one facility more than the other because of healthcare policy regulations. In fact, not all patients with cardiac and pulmonary diagnosis are sent to IRFs and SNFs. With missing values in the present data, the number of participants is naturally reduced further and becomes different between two facilities in the analysis of each matched dataset. For example, the participants were reduced from 27 to 18 for SNFs and 27 to 21 for IRFs in Fisher’s 1:1 Matched Data.

However, there is a dilemma between the simulation of the “real life” dataset and the limitation of analyzing data with small sample size due to missing values according to the Central Limit Theorem. Therefore, two recommendations can be given. First, when the number of participants can be maintained at a number greater than 30 for all levels of different groups in the analysis of each dataset, it is worthwhile utilizing no imputation procedure in order to simulate the “real life” situation. Second, when the number of participants is reduced to less than 30 in any level of different groups, the imputation should be utilized in order to improve the approximation of the analysis.

In two paragraphs above, the problem of missing values is clearly apparent, especially in data with small sample size in this study. The 33% of missing values in the posttest score of SF-12 in SNFs caused the approximation of the analysis unreliable due to lack of power in 1:1 Matched Data. This implies the clinicians—assigned study nurses or therapist in SNFs—who participated in this study did not collect data in a rigorous fashion like a researcher who is more aware of the serious research implications for missing data on analysis. If the clinicians recognized the importance of data collection, they could have followed up the patients even after the discharge, which requires only one phone call. In addition, according to the protocol of the formal study they were allowed to have a few follow-up phone calls. Therefore, it is necessary to educate clinicians the importance of collecting accurate data before the study starts. After comparing between central and local data of surgical performance thoroughly, Milburn et al. (2007, p. 275) stated, “The promulgation of inaccurate information could threaten reputation or career and clinicians should play a more active role in ensuring clinical data are correct.”

Based on the discussion, conclusions, and study implications and above, the present study’s recommendations can be summarized in a few points. First, this study recommends using not only RCM method but also Fisher’s method in matching with two considerations:1) in Fisher’s hand-matching, two matching methods are available using CVs—one with priority and another without; and 2) in prioritizing CVs, it is arguable which CVs should have higher priority. This study recommends that the prioritizing of CVs should be done in consultation with at least two experts or a general consensus in the

context of a study (e.g., health and rehabilitation in this present study). Otherwise, the results of the analysis of a matched dataset may be flawed due to matching participants through less-important background characteristics. The strength of prioritizing CVs is that more-important CVs are accounted for in matching comparable data. In RCM matching, many researchers have started prioritizing CVs by weighting them when calculating the propensity scores (Hirano & Imbens, 2001; Lunceford & Davidian, 2004; Rubin & Thomas, 1996; Robins and Rotnitzky ,1995), though Schafer and Kang (2008) discouraged this practice. Second, it is recommended to impute missing values to improve the approximation, especially when the sample size is small, less than 30, for all levels of different groups in the analysis of each matched data. Third, a caliper matching method is more useful with a small sample size data because it increases the power by adding more participants. Fourth, higher number of the same participants could be found between datasets made through the two matching methods, when CVs are treated in the same way, either with a priority or not, in both methods. Yet, when there are many CVs, the priority among CVs in hand-matching is still recommended due to "the curse of multi-dimensionality."

Limitations

Although utilizing RCM matching method with small-sample-size data is a new approach in this study, it has limitations. First, the number of participants in SNFs before matching is too small of a sample to represent the subjects. Therefore, the analysis of the data lacks generalizability. Second, after matching the data, the sample size was reduced

further from 27 to 18, and consequently lacks the power of the analysis due to many missing values in one of the outcome variables, SF-12D. Third, though both matching methods, Fisher's and RCM, posit causality between the treatment and the outcome measure, the causality cannot be established with the present data due to these two points: 1) based on Fisher's theory, the formal study was not executed with the experiment design with randomization and manipulation of the treatment; 2) it is also questionable to hold the assumption of "strongly ignorable treatment effect" in RCM theory because this study utilized a part of the whole data, only ten CVs, to match between the two groups. Fourth, most importantly, the datasets between two different matching techniques produced conflicting results on the main effect due to the increase of the sample size in one facility. Fifth, the correlation between the two DVs was low, 0.12-0.26, which negatively affected the power of the MANCOVA model. Tabachnick and Fidell (2007) stated that MANOVA has disadvantage with a low correlation between the DVs over separate ANOVA models in terms of statistical power. Sixth, and scales of the two DVs, SF-12 and of FIM, are different. Tabachnick and Fidell (2007) recommended using the same scale of measures as DVs in order to have best approximation with equal effect from several outcome measures in MANOVA. In order to compensate for the different scales of measures between the two DVs, it is useful to standardize both measures before the analysis, translating each value into Z-scores. Using Z-scores, however, requires normality, linearity and homogeneity of variance assumptions, which was not feasible in both Caliper Matched Data—Fisher's and RCM.

Future Studies

This present study proposes many areas to further intervene methodologically in matching the data. First, the present study can investigate further which method, Fisher's or RCM, produced a better matched dataset. Second, a few different imputation techniques can be utilized before matching. In this present study, no imputation methods were utilized mainly because any imputation procedure would produce incomparable datasets by applying it to only one of the two matched datasets between Fisher's and RCM. Also, leaving missing values resembles a real life situation. But, it may not be the case when an imputation technique is applied to both matching methods. Therefore, the present data could be investigated further with a few different imputation techniques. Third, it would also be interesting to compare datasets made through the two different methods, Fisher's method with priority of CVs and RCM method with weighted CVs. Additionally, it would be worthwhile to compare matched and unmatched datasets in IRFs. Finally, a discriminant analysis could be utilized to investigate membership criteria between the two facilities with the conditions of 1) facility differences in CVs and 2) more participants in SNFs. The discriminate analysis, however, is not possible with this present data due to limited number of participants in SNFs.

REFERENCE LIST

Abraham, B. & Ledolter, J., (2006). *Introduction to Regression Modeling*, (2nd ed). Duxbury Applied Series, Thomson Brooks.

Agresti, A., (2007). *An Introduction to Categorical Data Analysis* (2nd ed). Wiley series in probability and statistics, A John Wiley & Sons.

Albert, J.H. & Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 88*, 669-679

Box, J. F (1987). Guinness, Gosset, Fisher, and Small Samples. *Statistical Science, v2* n1, 45-52.

Beacham, A. (2008). An investigation of cardiac and pulmonary patient outcomes in inpatient rehabilitation versus skilled nursing facilities, preliminary project outcomes phase I, submitted to American Medical Rehabilitation Providers Association.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.M. (1984) *Classification and Regression Trees*. Boca Raton, FL: CRC Press

Cochran, W. G. (1954), The combination of Estimates from Different Experiments. *Biometrics, v10*, n1, 101-129.

Cochran, W. G. (1968), The effectiveness of adjustment by subclassification in Removing bias in observational studies, *Biometrics, 24*, 205-213.

Cochran, W. G., Rubin, D. B. (1974), Controlling bias in observational studies: A review. Mahalanobis Memorial Volume Sankhya –A, 1- 30.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver & Boyd.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of Ministry of Agriculture, 33*, 503-513.

Fisher, R. A. (1959). *Statistical Methods and Scientific Inference.* Edinburgh: Oliver and Boyd.

Fisher, R. A. and Mackenzie, W. A. (1923) [CP32]. Studies in crop variation, II: The manorial response of different potato varieties. *Journal of Agricultural Science*, Cambridge *13*, 311-320.

Fisher, R. A. and Wishart, J. (1930) [CP 85]. The Arrangement of Field Experiments and the Statistical Reduction of the Results. Technical Communication No. 10. Harperden, Hertfordshire: Imperial Bureau of Soil Science, Rothamsted Experimental Station.

Hansen, B. B. (2004). Full matching in an observational study of coaching for SAT. *Journal of the American Statistical Association*, *99*, 609- 619.

Hays, W. L. (1994). Statistics, 5th edition, Wadsworth, Thomson Learning, CA.

Ho, D., Imai, K., King, G., and Stuart, E. (2007), "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," *Political Analysis, 15*,199–236, http://gking.harvard.edu/files/abs/matchp-abs.shtml.

Holland, P. (1986). Statistics and casual inference. *Journal of the American Statistical Association*, *81*, 945-970.

Holland, P. W., & Rubin, D. B. (1980). Causal Inference in Prospective and Restrospective Studies. Address given at the Jerome Cornfield Memorial Session of the American Statistical Association Annual Meeting, August.

Hosmer, D.W. & Lemeshow, S. (1983). *Applied Logistic Regression* (2nd Ed). New York: Wiley

Hurlbert, S. H. (1984), Pseudoreplication and the design of ecological field experiments. *Ecological Monographs 54*, 187-211.

Kang, JD.Y. & Schafer, J.L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating population means from incomplete data. *Statistical Science*, *26*, 523-539

King, G. and Zeng, L. (2002). Improving forecasts of state failure. *World Politics, 53*, 623-658

King, J., Horowitz, M., Kassam, A., Yonaz, H., and Roberts, M. (2005, March). The Short Form-12 and the measurement of health status in patients with cerebral aneurysms: performance, validity, and reliability. *Journal of Neurosurg, 102*, 489-495.

Kuehl, R. O. (1994), *Statistical Principles of Research Design and Analysis*, Duxbury Press, Belmont, CA.

Kuehl, R. O. (2000), *Statistical Principles of Research Design and Analysis*, Duxbury Press, Belmont, CA.

Levin, I. P. (1999). *Relating Statistics and Experimental Design: An Introduction, Series: Quantitative Applications in the Social Science*. Sage university Papers, Sage Publications

Liu, C. (2004) Robit regression: a simple robust alternative to logistic and probit regression. In Applied Bayesian Modeling and Causal Inference from Incomplete-Data perspectives, A. Gelman and X. L. Meng (Eds), 227-238, New York: Wiley.

Mackintosh, S. (2009, March). Functional independence measure. *Australian Journal of Physiotherapy, 55*(1), 65. Retrieved from Academic OneFile database. (A208452709)

Manly, B. F. J. (2007). *Randomization, Bootstrap and Monte Carlo Methods in Biology,* 3rd edition, Texts in Statistical Science, Chapman & Hall/CRC, Taylor & Francis Group, FL.

Milburn, J.; Driver, C.; Youngson, G.; King, P.; MacAulay, E.; Krukowski, Z. (2007). The accuracy of clinical data: A comparison between central and local data collection. The Royal Colleges of Surgeons of Edinburgh and Ireland, 5, 275-278.

Neyman (Splawa), Jerzy (1923). On the application of probability theory to agricultural experiments. Essay on Principles. Section 9. *Statistical Scicence*. Translated and edited by D. M. Dabrowska and T. P. Speed from the Polish original, which appeared in Roczniki Nauk Rolniczych Tom X (1923) 1-51 (Annals of Agricultural Sciences).

Olson, C. L., (1976). On choosing a test statistics in multivariate analysis of variance. *Psychological Bulletin, 83*(4), 579-586.

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society*, Series B (Methodological), 53, 597-610.

---------(2002). *Observational Studies* (2nd ed.). New York: Springer

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for casual effects. *Biometrika, 70,* 41-55.

Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of American Statistical Association, 79,* 516-524.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics, 29*(1), 159-183.

Rubin, D. B. (1974), Estimating casual effects of treatments in randomized and nonrandomized studies of Educational *Psychology, 66,* 688-701.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology, 2,* 169-188.

Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of educational and Behavioral Statistics, v29* n3, 343-367.

Rubin, D. B. & Thomas, N. (2000). Combininig propensity score matching with additional adjustment for prognostic covariates. *Journal of the American Statistical Association, 95,* 573-485.

Stuart, E. A. and Green, K. M. (2008). Using Full Matching to Estimate Causal Effects in Nonexperimental Studies: Examining the Relationship Between Adolescent Marijuana Use and Adult Outcomes, *American Psychological Association, 44*(2), 395-406.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using Multivariate Statistics,* 5th edition, Ellyn & Bacon, Pearson, MA

VITA

Gideon Bahn was born and raised in Seoul, South Korea. Before attending Loyola University Chicago, he attended the Northeastern Illinois University, Chicago IL, where he earned a Bachelor of Arts and Science in Mathematics and Secondary Education in 2002. From 1981 to 1989, he also attended the A-Jou University, Suwon South Korea, where he received a Bachelor of Arts in Business Administration.

While at Loyola, Gideon earned a Masters of Science in Mathematics/Statistics in 2008, and taught statistics courses at the School of Business from 2008 to 2009. He also won the Chicago Community Trust award in 2005 and Diversity Faculty in Illinois Fellowship from 2006 to 2010.

Gideon had been working as a math teacher in a middle school from 2002 to 2005 and consulting various places as an independent investigator since 2007. Currently, he is a Biostatistician at Hines Veterans Affairs Medical Center in Hines, Illinois. He lives in Wheaton, IL.

DISSERTATION APPROVAL SHEET

The dissertation submitted by Gideon D. Bahn has been read and approved by the following committee:

Theresa Pigott, Ph.D., Director
Associate Professor of Research Methodology
Loyola University Chicago

Meng-Jia Bohanon, Ph.D.
Assistant Professor of Research Methodology
Loyola University Chicago

Kathleen Ruroede, Ph.D.
Assistant Vice President, Quality and Research
Marianjoy Wheaton Franciscan Healthcare

The final copies have been examined by the director of the dissertation and the signature which appears below verifies the fact that any necessary changes have been incorporated and that the dissertation is now given final approval by the committee with reference to content and form.

The dissertation is therefore accepted in fulfillment of the requirements for the degree of Doctor of Philosophy.

_____ _____

Date Director's Signature